Konstruktion computerisierter adaptiver Tests am Beispiel der Messung schulisch erworbener Kompetenzen

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena

von Raphael Bernhardt M. A.
geboren am 12.05.1985 **in** Mühlhausen

Gutachter

1. Prof. Dr. Andreas Frey, Friedrich-Schiller-Universität Jena

2. Prof. Dr. Susan Seeber, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 25.01.2017

1. Einleitung

1.1 Problemstellung und Argumentation

Die steigende Anzahl an Testpersonen, eine hohe Objektivität der Testung, effektives und schnelles Testen, sofortige Auswertung der Testergebnisse, unverzügliche Rückmeldung der Ergebnisse nach Testende, multimediale Elemente im Test und andere Anforderungen im Bildungsbereich haben dazu geführt, dass computerbasiertes Testen entwickelt und verwendet wurde und wird. Lerntheorien mit Blick auf personalisiertes (Studierenden-zentriertes) Lernen verlangten zudem individualisierte Testungen zur Überprüfung von Fähigkeiten bzw. Leistungen, unter anderem um die Probanden zu fordern, aber nicht zu frustrieren und um den Testverlauf sofort und kontinuierlich am Wissen, der Fähigkeit bzw. der Leistung des Probanden auszurichten. *Computerisiertes adaptives Testen* (CAT) zur Messung individueller Personenmerkmale (Frey, 2012) wird deshalb verstärkt angewandt. Beispielsweise wird der bekannte Englischtest TOEFL (*Test of English as a Foreign Language*) auch als computerisierter adaptiver Test angeboten (Economides & Roupas, 2007). CAT wird zumeist auf Basis der *Item Response Theorie*, kurz IRT, verwendet (Embretson & Reise, 2000). In Zukunft ist zu erwarten, dass sich Untersuchungen sowohl bei groß angelegten Vergleichsstudien als auch in der Individualdiagnostik dieser Testform annehmen. In der beruflichen Bildung gibt es bereits Bestrebungen computerbasierte Kompetenzmessungen durchzuführen, um die berufliche Handlungsfähigkeit von Auszubildenden zu messen. Die Ergebnisse sollen Informationen über die Leistungsstärken und -schwächen der Probanden liefern, um Bildungsprozesse optimieren zu können. In Deutschland wurden aufgrund der Debatte über groß angelegte Vergleichsstudien im Bereich der beruflichen Aus- und Weiterbildung (Achtenhagen & Baethge, 2008) diese Ziele im Forschungsprogramm *Technology-based Assessment of Skills and Competences in Vocational education and training* (*Technologie-orientierte Kompetenzmessung in der Berufsbildung*; ASCOT) verfolgt (Beck, Landenberger & Oser, 2016). Alle Projekte im Programm ASCOT arbeiteten mit modernen computerbasierten Verfahren, um Leistungsniveaus der Auszubildenden sichtbar zu machen. Ein Querschnittsprojekt der ASCOT Forschungsinitiative mit dem Namen *Messung allgemeiner Kompetenzen – adaptiv* (MaK-adapt) hatte die Aufgabe,

Instrumente zur Erfassung von Erklärungsvariablen beruflicher Fachkompetenz für alle ASCOT-Projekte bereitzustellen (Ziegler, Frey, Seeber, Balkenhol & Bernhardt, 2016). Dort wurden computerisierte adaptive Tests zur Messung der Kompetenzen von themenunabhängigen Grundqualifikationen (mathematische Kompetenzen, naturwissenschaftliche Kompetenzen und Lesekompetenz; nachfolgend auch schulisch erworbene Kompetenzen genannt) konstruiert, welche auf Schülerinnen und Schüler (SuS) beruflicher Schulen angewandt werden können. Zu Projektbeginn von MaK-adapt gab es deshalb folgende neuen Herausforderungen:

- Im berufsbildenden Bereich gibt es keine standardisierte Messung zur Ermittlung schulisch erworbener Kompetenzen. Es wird somit ein Test für ein neues Feld von Probanden entworfen.
- In den zu untersuchenden Kompetenzbereichen ist mit einer relativ breiten Streuung der Kompetenzausprägungen zu rechnen. Diese Annahme ergibt sich aus der Überlegung, dass die Zusammensetzung der Schülerinnen und Schüler in Berufsschulklassen bezüglich schulisch erworbener Kompetenzen aufgrund unterschiedlicher Vorerfahrungen, unterschiedlicher Herkunft (z. B. Schulen, Bundesland, soziokultureller Hintergrund) und unterschiedlicher Berufsabschlüsse sehr heterogen ist. Es wird somit ein Instrument benötigt, welches auch in den Randbereichen der möglichen Kompetenzausprägung zuverlässig misst.
- Aufgrund der Erhebung eigener Hauptstudien der anderen ASCOT-Projekte neben der Tests aus dem Projekt MaK-adapt und der damit verbundenen geringen Testzeit wurden für eine reliable Messung der schulisch erworbenen Kompetenzen Instrumente mit hoher Messeffizienz benötigt.

Eine angemessene Lösung für diese Herausforderungen bildet CAT. Wie oben beschrieben, ermöglicht CAT die Messung mehrerer Kompetenzen in geringerer Testzeit gleichbedeutend mit einer geringeren Belastung der Probanden und eine präzisere Messung der Kompetenzausprägung in den Randbereichen der Skala im Vergleich zu einem Test mit fester Itemreihenfolge (*Fixed Item Testing*; FIT). Deshalb war es naheliegend, in dem Anwendungsbereich des Forschungsprogramms ASCOT, CAT zu verwenden. Die Erstellung eines computerisierten adaptiven Tests benötigt im Unterschied zu üblich verwendeten Testformen wie papierbasiertem FIT meist mehr Ressourcen. Dass kann

z. B. zu höheren Kosten, höheren Entwicklungsaufwand, zusätzlich benötigter psychometrischer Expertise oder hohem Aufwand bei der Bereitstellung von Computern am Testort führen (Frey, 2012; Thompson & Weiss, 2011). Für die Entwicklung eines Itempools beim adaptiven Testen müssen beispielsweise viele Aufgaben in den unterschiedlichen Schwierigkeitsbereichen neu entwickelt werden, da hier selten ein normalverteilter, sondern ein gleichverteilter Itempool hinsichtlich der Itemparameter nützlich ist. Die Entwicklung von Items ist jedoch kosten- und zeitintensiv. Zudem ist die praktische Erstellung eines computerisierten adaptiven Tests methodisch anspruchsvoll. Wise und Kingsbury (2000) schreiben:

> The basic principles of computerized adaptive testing are relatively straightforward. The practice of implementing and maintaining an adaptive testing program, however, is far more complex. A number of practical challenges await measurement professionals […]. The success of an adaptive testing program will largely depend on how well the measurement practitioner deals with these challenges. (S. 135)

Für die Konstruktion der Tests im Projekt MaK-adapt waren relativ wenig Zeit und geringe finanzielle Ressourcen verfügbar. Zudem gab es für den Entwicklungsprozess von computerisierten adaptiven Tests bis auf einen Ansatz von Thompson und Weiss (2011) und den Hinweisen zu praktischen Problemen von Wise und Kingsbury (2000) keine dem Autor bekannten praktischen Anleitungen.

1.2 Ziele der Arbeit und Abgrenzung

Diese Arbeit stellt neben den theoretischen Grundlagen eine praktische Anleitung zur Entwicklung, Konstruktion und Administration eines computerisierten adaptiven Tests dar. Zudem werden die einzeln vorgestellten Schritte empirisch geprüft. Obwohl fast jede Testug in der Anwendung unterschiedlich und einzigartig ausfällt und aus architektonischer Sicht die Entwicklung eines adaptiven Tests als Baukastenprinzip mit den Bausteinen Itempool, Startpunkt, Itemauswahl usw. angesehen werden kann, wird diese Arbeit erstmals einen umfangreichen praktischen Rahmen zur Erstellung eines computerisierten adaptiven Tests darstellen. In dieser Arbeit wird beispielhaft gezeigt, wie mit geringen Mitteln in kurzer Zeit die notwendigen Schritte zur Entwicklung eines compu-

terbasierten adaptiven Tests in einem neuen heterogenen Feld durchgeführt werden können. Bei der Entwicklung der adaptiven Tests in dem hier verwendeten empirischen Beispiel MaK-adapt ergaben sich zwei zusätzliche Herausforderungen, welche durch weitergehende Zusatzstudien bearbeitet wurden und als grundsätzlicher Teil in den praktischen Rahmen mit einfließen.

Die erste zusätzliche Herausforderung im Projekt MaK-adapt war, Itempositionseffekte zu ermitteln und ggf. berücksichtigen zu können, da diese die IRT-Annahme der Invarianz der Itemparameterschätzungen verletzen können. Aktuelle Studien legen prinzipiell nahe, bei der Schätzung von Itemparametern Positionseffekte zu berücksichtigen (Albano, 2013; Debeer & Janssen, 2013; Hartig & Buchholz, 2012). Beim computerisierten adaptiven Testen werden zwar üblicherweise Annahmen zur Invarianz der Itemparameterschätzung für unterschiedliche Testsituationen und unterschiedliche Personengruppen kontrolliert. Die Itemposition als Grund für die Verletzung der Invarianzannahme wurde im Kontext von CAT bisher jedoch nicht thematisiert (Frey, Bernhardt & Born, im Druck). Grundsätzlich ist beim Vorliegen von Itempositionseffekten eine suboptimale Itemauswahl und eine verzerrte Merkmalsschätzung zu erwarten. Ein adaptiver Test sollte deshalb bei vorliegenden Itempositionseffekten nicht ohne weiteres angewandt werden. Aus diesem Grund wird nachfolgend ein mögliches Vorgehen gezeigt, Itempositionseffekte im Kontext von CAT zu ermitteln und damit umzugehen.

Die zweite Herausforderung stellte das Verbinden (*Linking*) eines computerisierten adaptiven Tests mit papierbasierten FIT dar. Durch die Entwicklung eines papierbasierten FIT konnte im Projekt MaK-adapt die Flexibilität des Einsatzes der Instrumente erhöht werden. Auf diese Weise kann auch in schwer zugänglichen Testfeldern (z. B. falls kein bzw. nicht ausreichend Computer vorhanden sind) erhoben werden. Der Einsatz zusätzlicher papierbasierter Tests mit fester Itemreihenfolge ist nur dann sinnvoll, wenn der FIT auf derselben Metrik wie der adaptive Test berichtet. Ein Linking setzt u. a. invariante Itemparameter über verschiedene Testformen voraus. Aufgrund von Faktoren, wie unterschiedliche Itempositionen, können Itemparameter zwischen Testformen variieren (Kolen & Brennan, 2014; Miller & Fitzpatrick, 2008). Da beim Testen mit fester Itemreihenfolge die Position, an der ein Item vorgelegt wird, konstant ist und beim computerisierten adaptiven Testen jedes Item an jeder Position auftauchen kann, ist

solch ein Linking nicht als Standardprozedur anzusehen. Deshalb wird nachfolgend eine Möglichkeit des Linking vorgestellt, welche mögliche Itempositionseffekte berücksichtigt. Die beiden Herausforderungen (Itempositionseffekte und Linking) können durch die vorliegende Arbeit in Zukunft zusätzlich bei der Entwicklung eines adaptiven Tests berücksichtigt und im Sinne einer beispielhaft vorgestellten Lösung bearbeitet werden. Das geschilderte Vorgehen eignet sich vor allem zur Erstellung von Testungen im Kompetenzbereich, in dem standardisierte, schnell auswertbare Single-Choice und Multiple-Choice-Items sowie kurze offene Items eingesetzt werden können. In dieser Arbeit wird sich ausschließlich auf computerisierte adaptive Tests im Rahmen der IRT bezogen, da die IRT die Berechnung gleicher Punktwerte über unterschiedliche Mengen vorgegebener Items erlaubt. Überlegungen zum adaptiven Testen ohne die IRT finden sich z. B. bei Yan, Lewis und Stocking (2002).

1.3 Aufbau dieser Arbeit

Diese Arbeit bietet eine Anleitung zur Erstellung eines computerisierten adaptiven Tests und dem Linking mit einem papierbasierten FIT in sechs Schritten. Diese sechs Schritte wurden in dieser Arbeit als theoretischer Rahmen vom Autor erarbeitet. Aus diesem Grund findet vor der eigentlichen praktischen und empirischen Prüfung eine umfangreiche theoretische Erläuterung der Inhalte dieser Schritte statt. Erste praktische Tätigkeiten z. B. zur Testplanung erfolgen erst ab Kapitel 4. Die erste empirische Studie befindet sich im Kapitel 4.3 (Kalibrierungsstudie). Die Unterkapitel im empirischen Teil sind größtenteils klassisch nach Fragestellungen, Methode, Ergebnisse und Zusammenfassung aufgebaut. Mit dem Thema CAT vertraute Leser können dank dieser Struktur direkt in den praktischen und empirischen Teil (Kapitel 4) übergehen und bei Bedarf die entsprechenden theoretischen Abhandlungen der einzelnen Schritte nachlesen.

Im nachfolgenden Kapitel 2 werden die übergreifenden Forschungsfragen dieser Arbeit aufgeführt. Anschließend werden allgemeine theoretische Grundlagen sowie die Anleitung zur Erstellung eines computerisierten adaptiven Tests erarbeitet. Da in dieser Arbeit nur CAT im Rahmen der IRT betrachtet wird, gibt es im Kapitel 3.1.1 eine Einführung in die Grundlagen der IRT. Das Kapitel 3.1.2 enthält das notwendige Grundlagenwissen über (computerisiertes) adaptives Testen, um die weiteren Kapitel besser

verstehen zu können. Lesern ohne Kenntnisse über die IRT und CAT wird empfohlen, Kapitel 3.1 vor den einzelnen Schritten der Anleitung zur Erstellung eines adaptiven Tests (Kapitel 3.2 bis Kapitel 3.7) zu lesen.

Die Schritte zur Entwicklung eines computerisierten adaptiven Tests beginnen ab Kapitel 3.2 mit der Testplanung. Hier wird u. a. die Festlegung des inhaltlichen Zielkonstrukts als ein wichtiger Bestandteil im Kapitel 3.2.1 näher erläutert. Weiterhin wird auf die Verwendung von Simulationsstudien als ein bedeutsames Werkzeug bei der Testplanung und -erstellung hingewiesen (vgl. Kapitel 3.2.2). Als weiterer Teil der Testplanung ist wird Im Kapitel 3.2.3 auf Fragen nach der geeigneten Software und der technischen Umsetzung eingegangen. Nachdem die vorläufigen Parameter durch die Simulationsstudien, die zu verwendende Software und das inhaltliche Zielkonstrukt festgelegt wurden, wird als zentraler Bestandteil des Tests ein Itempool benötigt. Im Kapitel 3.3 werden Schritte zur Erstellung eines Itempools aufgeführt. Dabei geht es vor allem um die speziellen Anforderungen eines Itempools im Zusammenhang mit computerisierten adaptiven Testen (vgl. Kapitel 3.3.1) und um die Entwicklung von Items spezifisch für CAT (vgl. Kapitel 3.3.2). Ein Itempool kann im Kontext für CAT erst verwendet werden, wenn er getestet und kalibriert wurde (vgl. Kapitel 3.4). Aus diesem Grund wird im Kapitel 3.4.1 auf die Grundlagen zur Durchführung einer Kalibrierungsstudie eingegangen. Das *Testheftdesign* spielt hierbei eine besondere Rolle, da es über die Qualität der Kalibrierung mitentscheidet. Die Kalibrierungsstudie wird primär dazu verwendet, die relevanten Itemparameter, die später im adaptiven Algorithmus benötigt werden, zu schätzen und die Itemqualität zu bestimmen bzw. wenig qualitative Items aus dem Itempool zu entfernen. Aus diesem Grund wird im Kapitel 3.4.2 auf Fragen zur Itemparameterschätzung, Itemqualität und zum Modellfit eingegangen. Ein weiterer Schritt zur Sicherstellung der Qualität des Itempools ist die Prüfung der Items auf Differential Item Functioning (DIF), was in Kapitel 3.4.3 beschrieben wird. Als zusätzlicher Schritt bei der Testung und Kalibrierung des Itempools wird hier der Schritt zur Prüfung von Itempositionseffekten (vgl. Kapitel 3.4.4) eingeführt. Neben dem Itempool benötigt CAT einen adaptiven Algorithmus. Im Kapitel 3.5 werden auf die wesentlichen Aspekte Startpunkt (vgl. Kapitel 3.5.1), Itemauswahl (vgl. Kapitel 3.5.2), Schätzung von Personenparametern (vgl. Kapitel 3.5.3), Beendigung des Tests (vgl. Kapitel 3.5.4) und Restriktionen an den Test (vgl. Kapitel 3.5.5) eingegangen. Das Zusammenbringen der Komponenten Item-

pool, Algorithmus und Software erfolgt in Kapitel 3.6, wo der Test veröffentlicht und in einer Pilotierungsstudie (vgl. Kapitel 3.6.1) angewendet wird. Zusätzlich werden in diesem Kapitel noch Hinweise zur Bildung (vgl. Kapitel 3.6.2) und zum Erhalt (vgl. Kapitel 3.6.3) einer Skala gegeben, wobei Aspekte wie Testsicherheit, Itemparameterdrift und das Hinzufügen bzw. Entfernen von Items eine Rolle spielen. Im Kapitel 3.7 wird letztendlich darauf eingegangen, wie ein computerisierter adaptiver Test mit anderen Testarten verbunden bzw. gleichgesetzt werden kann.

Kapitel 4 liefert die empirischen Befunde und zeigt eine praktische Anwendung der in Kapitel 3 aufgeführten theoretischen Schritte. Dabei sind die Unterkapitel so aufgebaut, dass sie entsprechend einer empirischen Arbeit stets Fragestellungen, Methode und Ergebnisse aufweisen. Das Kapitel 4.1 bezieht sich auf die Testplanung und beschreibt vor allem die Inhalte und Ziele des Projekts MaK-adapt (vgl. Kapitel 4.1.2). Als praktische Anwendung der vorher aufgeführten Schritte wird hier auf die Festlegung des inhaltlichen Zielkonstrukts (vgl. Kapitel 4.1.3) sowie die softwaretechnische Umsetzung (vgl. Kapitel 4.1.4) eingegangen. Bei der Entwicklung des initialen Itempools (vgl. Kapitel 4.2) wurde im Projekt MaK-adapt eine Methode verwendet, die hier als Itemrecycling umschrieben wird (vgl. Kapitel 4.2.2). Zudem wird auf die Computerisierung der Items mit der verwendeten Software *Multidimensional Adaptive Testing Environment* (MATE) eingegangen.

Im Kapitel 4.3 geht es um den Pretest und die Kalibrierung des Itempools. Neben dem Ablauf der Kalibrierungsstudie und der Stichprobenbeschreibung (vgl. Kapitel 4.3.3) werden die Methoden und Ergebnisse der Kalibrierungsstudie (Itemparameterschätzung, Itemselektion und DIF-Analysen) beschrieben. Besonderes Augenmerk ist bei der Kalibrierungsstudie auf das Testheftdesign (vgl. Kapitel 4.3.2) gelegt worden. Zudem wird die vorgeschlagene Methode zur Überprüfung von Itempositionseffekten an empirischen Daten erprobt (vgl. Kapitel 4.3.5). Kapitel 4.4 zeigt, wie die einzelnen Schritte des computerisierten adaptiven Algorithmus entsprechend des Pfaddiagramms zum Ablauf computerisierter adaptiver Tests (vgl. Abbildung 2 auf S. 59) im vorliegenden empirischen Fall spezifiziert werden können. Im Kapitel 4.5 wird auf die Veröffentlichung und Anwendung des computerisierten adaptiven Tests eingegangen. Neben der Stichprobenbeschreibung der Pilotierungsstudie für CAT (vgl. Kapitel 4.5.2) werden die Methode und Ergebnisse der Pilotierungsstudie vorgestellt und Anpassungen für den

Algorithmus sowie den Itempool abgeleitet (vgl. Kapitel 4.5.3). Kapitel 4.6 beschreibt, wie die Skala eines papierbasierten Tests mit fester Itemreihenfolge und die Skala eines computerisierten adaptiven Tests verbunden werden können. Dazu wird zuvor der Ablauf (vgl. Kapitel 4.6.2) und die Stichprobe (vgl. Kapitel 4.6.3) der Pilotierungsstudie der papierbasierten Testversion beschrieben. Zudem wird die Methode des Linking vorgestellt (vgl. Kapitel 4.6.4) und empirisch geprüft (vgl. Kapitel 4.6.5).

Im Kapitel 5 werden die einzelnen Schritte zur Testerstellung sowie die empirischen Ergebnisse dazu zusammenfassend diskutiert sowie auf den praktischen Beitrag dieser Arbeit eingegangen. Anschließend wird Ausblick gegeben und ein Fazit gezogen.

2. Fragestellungen

Aufgrund der geschilderten Problemstellung aus Kapitel 1.1 wurden vier Hauptfragen hergeleitet und anschließend in dieser Arbeit beantwortet. Diese vier Fragen werden nachfolgend mit einer kurzen Erläuterung aufgeführt. Aufgabe des Projektes MaK-adapt war es, effiziente Messinstrumente für die drei Domänen Lesen, Mathematik und Naturwissenschaft, die auch in den Randbereichen der Kompetenzverteilung bei einer heterogenen Stichprobe angemessen differenzieren, zu entwickeln. Die Entwicklung der Testinstrumente musste in relativ kurzer Zeit (max. 18 Monate) erfolgen, damit diese von den ASCOT-Projekten in deren Haupterhebung genutzt werden konnten (Ziegler et al., 2016). Die Lösung für diese Herausforderung stellte CAT dar. Da das Erstellen von computerisierten adaptiven Tests in der Regel zeitaufwendiger, methodisch anspruchsvoller und teurer ist als die Erstellung eines papierhaften Tests mit fester Itemreihenfolge (Frey, 2012), bildet die erste Fragestellung einen zentralen Aspekt dieser Arbeit.

1) Wie lässt sich ein computerisierter adaptiver Test zur Messung schulisch erworbener Kompetenzen in einem neuen heterogenen Feld mit geringen finanziellen und zeitlichen Ressourcen verwirklichen?

Zu Beginn der ASCOT-Initiative lagen weder effiziente Messinstrumente zur Messung schulisch erworbener Kompetenzen bei SuS beruflicher Schulen (Ziegler et al., 2016) noch umfangreiche praktische Anleitungen zur Erstellung computerisierter adaptiver Tests vor (Thompson & Weiss, 2011). Zudem war der aktuellste dem Autor bekannte praktische Rahmen von Thompson und Weiss (2011) als Zeitschriftenartikel recht kurz gehalten und deckte nicht die aktuellen Anforderungen (die Erstellung der Tests mit geringen Ressource in kurzer Zeit, die Betrachtung von Itempositionseffekten und das Linking mit FIT) ab. Die Anforderung, Itempositionseffekte bei der Schätzung von Parametern zu berücksichtigen, ist so bedeutsam, da diese Effekte Annahmen der IRT verletzen und somit Item- und Personenparameter verzerren können (Albano, 2013). Da mögliche Itempositionseffekte beim computerisierten adaptiven Testen nicht etwa durch Testhefte statistisch kontrolliert werden können (Frey, Hartig & Rupp, 2009), bezieht sich die zweite Frage auf ein wichtiges Element zur Erstellung computerisierter adaptiver Tests.

2) Wie lassen sich Positionseffekte bei der Entwicklung eines adaptiven Tests angemessen ermitteln und ggf. berücksichtigen?

Doch verzerrte Item- und Personenparameter haben nicht erst Einfluss bei der Itemauswahl und Personenparameterschätzung im adaptiven Algorithmus. Bereits bei der Kalibrierungsstudie sind Effekte auf die Personenverteilung und die Itemselektion zu erwarten. Die dritte Fragestellung erweitert deshalb den in der zweiten Fragestellung angeknüpften Punkt.

3) Welche Relevanz hat die Berücksichtigung von Positionseffekten auf die Personenverteilung und die Itemselektion der MaK-adapt-Kalibrierungsstudie?

Mit der dritten Fragestellung wird auch verdeutlicht, dass Itempositionseffekte gerade in Bezug auf CAT, wo über ein fixes Testheftdesign fixe Itemparameter ermittelt, aber später im adaptiven Test die Items flexibel vorgegeben werden, eine praktische Relevanz besitzen. Die Fragen zu Itempositionseffekten werden in Werken zum Thema CAT bisher nicht aufgegriffen (Frey et al., im Druck). Die zweite und dritte Fragestellung gehen daher auf den bisher eher vernachlässigten Bereich der Itempositionseffekte bei der Testentwicklung ein. Ziel ist es, die Forschungslücke zu diesem Thema ein Stück weit zu schließen und die Ergebnisse in Form einer Standardprozedur zur Berücksichtigung von Positionseffekten mit in den Rahmen der Testentwicklung einfließen zu lassen.

Eine weitere Anforderung an den praktischen Rahmen stellt das Linking (Kolen & Brennan, 2014) eines computerisierten adaptiven Tests an einen papierbasierten Test mit fester Itemreihenfolge dar. Aufgrund der unterschiedlichen Einsatzbereiche von Testinstrumenten ist es nicht immer möglich oder erwünscht, ausschließlich computerbasiert zu testen. Sollen beispielsweise vorhandene papierbasierte Testinstrumente parallel laufen oder wie im Projekt MaK-adapt durch FIT die Einsatzmöglichkeiten erhöht werden, ist es wichtig, dass beide Testarten auf derselben Metrik berichten. Für das Projekt MaK-adapt war es eine Herausforderung, mit den unterschiedlichen technischen Ausstattungen der Schulen umzugehen und die Software zur Administration der adaptiven Tests ohne Probleme ausführen zu können (z. B. fehlende Computer-Arbeitsplätze, keine Administratorrechte für die Installation der Tests, fehlende Zusatz-Software, kein Internetzugang). Eine Möglichkeit, SuS aus solchen Schulen dennoch

testen zu können, stellt papierbasiertes FIT dar (Ziegler et al., 2016). Daraus ergibt sich die vierte Fragestellung:

4) Wie lassen sich unter den Bedingungen vom Projekt MaK-adapt die Skala eines papierbasierten Tests mit fester Itemreihenfolge und die Skala eines computerisierten adaptiven Tests angemessen miteinander verbinden?

Bisher wird das Linking eines papierbasierten Tests mit fester Itemreihenfolge mit einem computerisierten adaptiven Test von der Literatur nicht als Schritt zur Erstellung eines computerisierten adaptiven Tests gesehen. In dieser Arbeit wird dies als optionaler letzter Schritt eingeführt, um standardmäßig eine Schnittstelle zum FIT zu erhalten.

3. Theoretische Grundlagen

Dieses Kapitel zeigt theoretisch die notwendigen Schritte zur Erstellung eines computerisierten adaptiven Tests. Der hier vorgestellte theoretische Rahmen orientiert sich teilweise an dem vorgeschlagenen Rahmen zur Entwicklung eines adaptiven Tests von Thompson und Weiss (2011), die ein allgemeingültiges und zugleich spezifisches Modell für den Testentwicklungsprozess im Rahmen eines computerisierten adaptiven Tests aufstellen. Die vorgeschlagenen fünf Schritte von Thompson und Weiss finden sich hier in abgewandelter und erweiterter Form wieder.

Schritt 1: Die Durchführbarkeit, die Anwendbarkeit und die Planung von Studien befinden sich im Kapitel 3.2 (Testplanung). Dieses Kapitel wurde erweitert durch die Festlegung des inhaltlichen Zielkonstrukts, der Software und der technischen Umsetzung sowie der Nutzung von Simulationsstudien.

Schritt 2: Die Entwicklung des initialen Itempools wurde im Kapitel 3.3 untergebracht. Dabei werden Anforderungen des Itempools besprochen und Hinweise zur Entwicklung von Items gegeben.

Schritt 3: Der Pretest und die Kalibrierung des Itempools finden sich im Kapitel 3.4 (Pretest und Kalibrierung des Itempools) wieder.

Schritt 4: Die Festlegung der Spezifikationen für den finalen computerisierten adaptiven Test wurde im Kapitel 3.5 (CAT – Algorithmus) berücksichtigt.

Schritt 5: Die Veröffentlichung des computerisierten adaptiven Tests ist im Kapitel 3.6 (CAT – Veröffentlichung und Anwendung) enthalten.

Mit dem Kapitel 3.7 (Linking mit papierbasierter Testung) ist ein weiterer Schritt hinzugekommen, in welchem die Verbindung von Skalen bei der Nutzung von zwei oder mehreren Testarten (z. B. papierbasierte Testung oder computerisierte Testung mit fixer Itemreihenfolge) behandelt wird. Zuvor wird im Kapitel 3.1 ein Überblick über die Grundlagen der Item Response Theorie und des computerisierten adaptiven Testens gegeben.

3.1 Item Response Theorie (IRT) und computerisiertes adaptives Testen (CAT)

Dieser Abschnitt gibt einen knappen Einstieg in die Grundlagen der IRT und des computerisierten adaptiven Testens. Zusätzlich wird in diesem Abschnitt auf Aspekte der Reliabilität, der Validität und der Motivation im Zusammenhang mit adaptivem Testen eingegangen. Vor allem Testentwicklern ohne Vorkenntnisse in den Bereichen IRT und CAT wird empfohlen, sich das Kapitel 3.1 vor dem Weiterlesen anzusehen und ggf. weitere Grundlagenliteratur hinzuzuziehen.

3.1.1 Grundlagen der IRT

Der Fokus der IRT liegt auf der Antwort einer Person auf ein Item. Diese Antwort wird als Resultat eines Zufallsprozesses modelliert, in der das Antwortverhalten der Probanden durch ein mathematisches (probabilistisches) Modell abgebildet wird. Dabei hängt die Wahrscheinlichkeit einer korrekten Antwort von unterschiedlichen Parametern ab, typischerweise den Itemparametern und den Personenparametern. Zusätzlich können dem Modell weitere Parameter hinzugefügt und auch Interaktionsparameter z. B. für die Interaktion der Personen mit den Items verwendet werden. Die Personenparameter bilden üblicherweise die Fähigkeit oder das Wissen einer Person in einem bestimmten Bereich (z. B. Mathematikwissen) ab und werden auch als Fähigkeitsparameter bezeichnet. Die Itemparameter stehen häufig für die Itemdiskrimination a, die Itemschwierigkeit b und die Ratewahrscheinlichkeit c. In der vorliegenden Arbeit wird die Eigenschaft einer Person, die mit den Items gemessen wird (der zu messende Trait), als Personen- bzw. Fähigkeitsparameter Theta (θ) bezeichnet. Dabei werden mathematische Funktionen (Item-Response-Funktionen, IRF) genutzt, um bei gegebenem θ die Wahrscheinlichkeit einer Person, ein Item korrekt zu beantworten, zu berechnen. Es wird demzufolge das Antworten auf ein Item als Wahrscheinlichkeitsfunktion der Personen- und Itemmerkmale modelliert (van der Linden & Hambleton, 2005). Item- und Personenparameter können so auf einer gemeinsamen Skala berichtet werden. Die Modelle der IRT können auf unterschiedlichste Weise inhaltlich unterteilt werden. Es kann zwischen ein- und mehrdimensionalen (bzw. multidimensionalen) Modellen, zwischen Modellen für dichotome und polytome (ordinale) Daten und zwischen Modellen mit unterschiedlicher Anzahl an Itemparametern (1PL, 2PL, 3PL) unterschieden werden. Nachfolgend wird hauptsächlich auf das unidimensionale 1PL-Modell für dichotome Daten eingegangen, da

die Entwicklung der adaptiven Tests im empirischen Teil auf Grundlage des 1PL- bzw. Rasch-Modells für dichotome Daten beruht. Das 2PL- und 3PL- Modell werden nur kurz aufgeführt. Alle genannten Testmodelle für dichotome Daten gibt es verallgemeinert auch für ordinale Daten. Dort werden die *Item Characteristic Curves* (ICCs) wesentlich komplexer. Ein Beispiel für ein ordinales Rasch-Modell findet sich z. B. bei (Rost, 2006).

Eindimensionale Modelle (unidimensionale IRT, UIRT) für dichotome Daten

Dichotome Items unterscheiden nur zwei Antwortkategorien (z. B. korrekt und falsch). Die korrekte Antwort wird beispielsweise mit 1 bewertet und die falsche Antwort mit 0. Für dichotome Modelle reicht es aus, sich auf die IRF der korrekten Antwort (der mit 1 gescorten Antwort) zu konzentrieren, da die Wahrscheinlichkeit für die falsche Antwort die Gegenwahrscheinlichkeit abbildet. Beide Antwortwahrscheinlichkeiten addieren sich zu 100 % bzw. zu 1. Die Antwortfunktion für die korrekte Antwort kann neben der IRF auch als ICC abgebildet werden (Rost, 2006). Die IRF bzw. die ICC beschreiben in der IRT eine nicht-lineare Beziehung zwischen der Wahrscheinlichkeit eines gezeigten Antwortverhaltens eines Probanden in Abhängigkeit von seiner Ausprägung auf dem zugrundeliegenden latenten Merkmal θ (Embretson & Reise, 2000).

Damit die IRT aus theoretischer Sicht gilt, müssen im unidimensionalen Fall folgende Annahmen erfüllt sein: Das zu messende individuelle Merkmal (Trait) bezieht sich im unidimensionalen Fall auf eine latente Merkmalsdimension und wird mit θ gekennzeichnet. Der wahre Wert von θ für eine Person verändert sich während der Testung nicht. Die Wahrscheinlichkeit einer korrekten Antwort auf ein Item kann auf genau eine latente Merkmalsdimension zurückgeführt werden. Der Zusammenhang zwischen der Wahrscheinlichkeit einer korrekten Antwort auf ein Item und der Ausprägung auf der latenten Merkmalsdimension kann mit einer kontinuierlich, monoton steigenden Funktion beschrieben werden (Monotonie). Die Charakteristika der Testitems sind unabhängig von der Testsituation. Die Antworten einer Person auf ein Item hängen nicht von vorhergehenden beantworteten Items ab (lokale stochastische Unabhängigkeit). Item- und Personenparameter sind stichprobenunabhängig. Der Standardmessfehler variiert in Abhängigkeit von der Ausprägung von θ (Embretson & Reise, 2000; Rost, 2006). Eine weitere häufig implizierte Annahme ist, dass die Testung nicht unter

Zeitdruck bzw. Geschwindigkeitsbedingungen (Speededness) erfolgt. Wenn Geschwindigkeitsbedingungen die Testleistung beeinflussen, dann müssen zwei Merkmale untersucht werden: zum einen die Geschwindigkeit und zum anderen das eigentlich zu messende Merkmal (z. B. Wissen). Vor einer Anwendung der IRT ist zu prüfen, ob die Testzeit einen Einfluss auf die Beantwortung der Items hat und wie viele Probanden den Test nicht vollständig bearbeitet haben (Hambleton & Swaminathan, 1985).

Häufig verwendete Testmodelle sind das Guttman-Skalenmodell, das Rasch-Modell, das Proctor-Modell, das Keats-Modell, das Latent-Distance-Modell, das Birnbaum-Modell, das Drei-Parameter logistische Modell, das Normalogiven-Modell oder das Binomial-Modell (Rost, 2006). In dieser Arbeit wird nur auf Modelle mit logistischer Linkfunktion eingegangen. Es kann zwischen Ein-Parameter logistischen (1PL, häufig auch als Rasch-Modell bezeichnet), Zwei-Parameter logistischen (2PL, häufig auch als Birnbaum-Modell bezeichnet) und Drei-Parameter logistischen (3PL) Modellen unterschieden werden.

Rasch-Modell (1PL-Modell)

Die mathematische Formel des 1PL- und des Rasch-Modells (Rasch, 1980) sind zwar nicht identisch, doch das 1PL-Modell ist rechnerisch eine einfache Annäherung an das Rasch-Modell (Lord, 1980). Numerisch und grafisch unterscheiden sich beide Modelle kaum (Rost, 2006). Deshalb wird hier weiterführend das Rasch-Modell mit folgender logistischer Formel verwendet und synonym als 1PL-Modell bezeichnet.

$$P(X_{ui} = 1) = \frac{e^{(\theta_u - b_i)}}{1 + e^{(\theta_u - b_i)}} \qquad (1)$$

Dabei ist $P(X_{ui} = 1)$ die Wahrscheinlichkeit für eine Person u das Item i korrekt zu beantworten. Die Fähigkeit einer Person u wird mit θ_u und die Schwierigkeit eines Items i mit b_i bezeichnet. Die Ausprägung $X_{ui} = 1$ bedeutet im Rahmen des dichotomen Antwortmodells, dass die Antwort von Person u auf das Item i korrekt war. Im Rasch-Modell spielt somit lediglich die Itemschwierigkeit eine Rolle als Itemparameter. Dies bedeutet für die ICCs, dass diese parallel zueinander verlaufen und denselben Anstieg haben. Die ICCs können sich somit beim Rasch-Modell nicht überschneiden. Nachfolgend

sind drei ICCs für das Rasch-Modell mit den Schwierigkeitsparametern $b = -0.051, b = 0.650$ und $b = 1.108$ abgebildet.

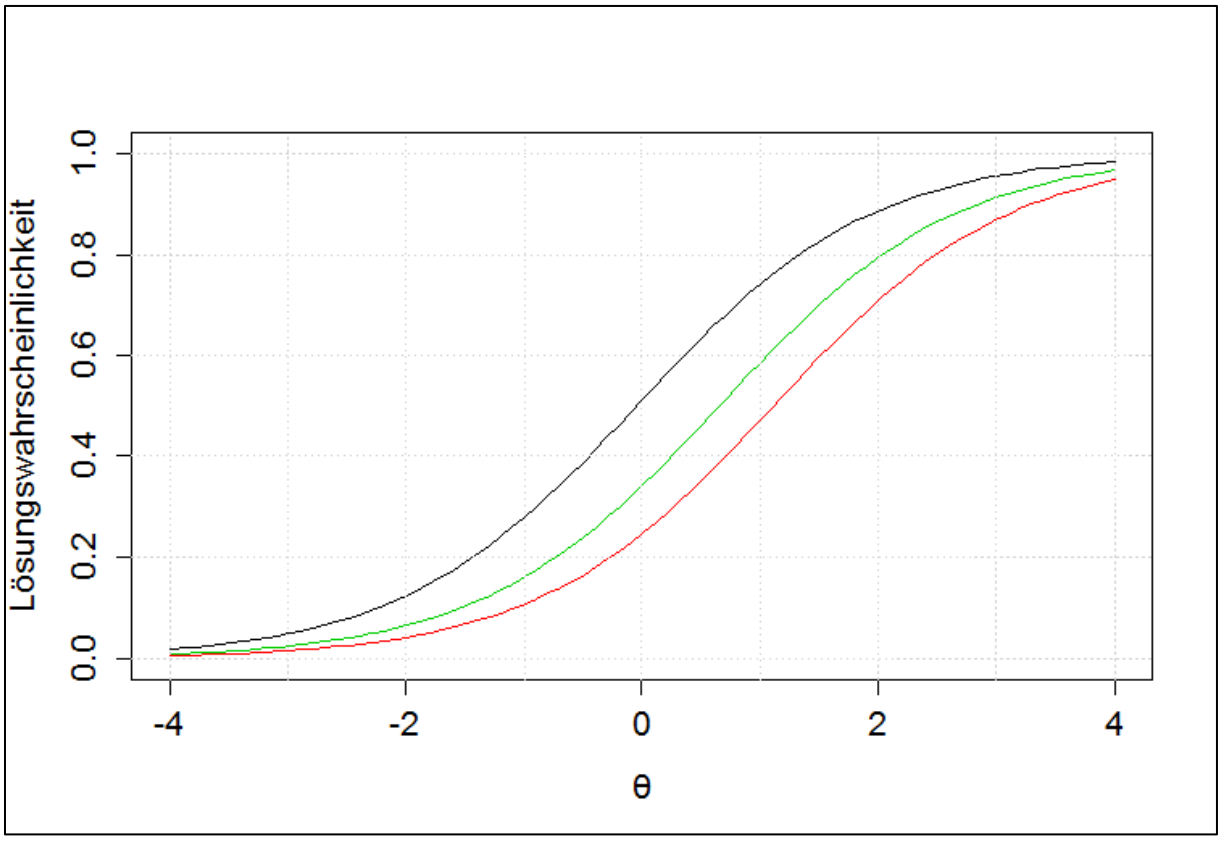

Abbildung 1: ICCs des Rasch-Modells für drei Items mit den Schwierigkeitsparametern b=-0.051, b=0.650 und b=1.108

Wichtig anzumerken ist, dass im Rasch-Modell die Annahmen gelten, dass in allen denkbaren Teilstichproben von Personen derselbe Schwierigkeitsparameter für ein Item geschätzt wird und dass die Personenparameter für alle möglichen Mengen von Items konstant sind (Rost, 2006).

Exkurs: Birnbaum-Modell (2PL-Modell) und 3PL-Modell

In einigen Anwendungsbereichen zeigt sich, dass das Rasch-Modell das Antwortverhalten auf die verwendeten Items nicht gut abbildet. Ein Grund für diese Nichtpassung, dem sogenannten *Misfit*, kann sein, dass beim Rasch-Modell die ICCs parallel verlaufen müssen. Wenn diese Eigenschaft nicht für alle Items zutrifft, ist eine Lösungsmöglichkeit, die unpassenden Items mit abweichendem Anstieg aus dem Itempool zu entfernen. Eine andere Möglichkeit ist, das Rasch-Modell zu generalisieren und unterschiedliche

Anstiege der ICCs zu erlauben (Wainer & Mislevy, 2000). Dies tut das 2PL-Modell, welches erstmals von Birnbaum (1968) beschrieben wurde. Im 2PL-Modell gibt der zusätzliche Parameter a_i den Anstieg und somit die Diskrimination des Items i an.

$$P(X_{ui} = 1) = \frac{e^{a_i(\theta_u - b_i)}}{1 + e^{a_i(\theta_u - b_i)}} \tag{2}$$

Das 3PL-Modell (Hambleton, 1982) enthält als Erweiterung des 2PL-Modells einen dritten Parameter, den sogenannten Pseudo-Rateparameter c_i, welcher die Ratewahrscheinlichkeit abbildet. Beim 3PL-Modell wird der Zusammenhang zwischen der Wahrscheinlichkeit, ein Item i mit der Diskrimination a_i, der Schwierigkeit b_i und dem Pseudo-Rateparameter c_i zu lösen, und der Merkmalsausprägung θ_u eines Individuums u durch folgende logistische Funktion beschrieben:

$$P(X_{ui} = 1) = c_i + (1 - c_i)\frac{e^{a_i(\theta_u - b_i)}}{1 + e^{a_i(\theta_u - b_i)}} \tag{3}$$

Exkurs: Multidimensionalität

Aus psychometrischer Sicht gibt es unterschiedliche Gründe, warum ein Test als multidimensional anzusehen ist (Reckase, 2009). Beispielsweise kann der Test mehrdimensional angelegt sein, indem mehrere latente Traits gemessen werden sollen. Um Abweichungen der Daten von der Annahme der Eindimensionalität zu entdecken, bietet sich z. B. die Nutzung der latenten Klassenanalyse an. Die Faktorenanalyse bietet eine weitere Möglichkeit, die Dimensionalität eines Tests zu ermitteln (Rost, 2006). Eine Alternative zur Modellierung von Multidimensionalität in Bezug auf verschiedene Teilpopulationen bietet das Mixed-Rasch-Modell (Rost, 1990). Hier gilt die Annahme der Eindimensionalität zwar innerhalb von Teilpopulationen, aber nicht für die Gesamtpopulation. Eine detaillierte Betrachtung multidimensionaler Modelle erfolgt an dieser Stelle nicht, da im weiteren Verlauf dieser Arbeit die Entwicklung eines unidimensionalen Tests beschrieben wird. Umfangreichere Ausführungen multidimensionaler Modelle finden sich u. a. bei Reckase (2009) und Rost (2004).

Modellgültigkeitskontrolle

Wie gut passt ein Modell zu den Daten? Das ist bei der Auswahl des Testmodells eine zentrale Frage, die auf unterschiedliche Weise beantwortet werden kann. Dabei ist zu berücksichtigen, dass eine gute Übereinstimmung mit den empirischen Daten nicht das einzige Ziel der Theoriebildung sein kann, sondern dies mit möglichst wenigen und einfachen Annahmen erreicht werden sollte. Die Theorie sollte deshalb dem Einfachheitskriterium folgen und empirische Gültigkeit aufweisen. Die Modellgültigkeitskontrolle prüft deshalb: (a) Wie gut erklärt das Modell die Daten? (b) Mit welchem Aufwand an Modellparametern wird diese Güte erreicht? und (c) Wie gut passt das Modell zum Forschungsstand? (Rost, 2004). Wenn bereits ein Testmodell bevorzugt wird, kann dessen Annahmen im Rahmen eines Modellgültigkeitstests geprüft und ggf. mit anderen konkurrierenden Modellen verglichen werden. Hier ist als Beispiel der Andersen-Test (Andersen, 1973) zu nennen, der das Rasch-Modell mit einem Mischverteilungsmodell vergleicht und prüft, ob die Itemparameter in verschiedenen Teilstichproben übereinstimmen. Der Martin-Löf-Test (Martin-Löf, 1977) entspricht einem Vergleich des eindimensionalen mit dem zweidimensionalen Rasch-Modell (zur Prüfung, ob die Personenparameter für alle denkbaren Untergruppen von Items konstant sind). Alle Modelltests lassen sich als Vergleich zwischen zwei konkurrierenden Modellen auffassen (Rost, 2004). Das Problem der modellvergleichenden Tests ist, dass meist die Voraussetzung für die Durchführung eines Signifikanztests nicht gegeben ist. Als Ausweg kann auf inferenzstatistische Schlüsse verzichtet werden. In diesem Fall wird sich auf informationstheoretische Maße beschränkt (vgl. Kapitel 3.4.2). Als Alternative kann mittels Bootstrapping Verfahren eine Prüfverteilung simuliert werden. Dabei werden wiederholt Statistiken auf Grundlage derselben Stichprobe berechnet (Stichprobenwiederholung), welche einen statistischen Schluss erlauben (Rost, 2004; Rost, 2006). Genaueres zum Modellgültigkeitstest findet sich bei der Beschreibung der Kalibrierungsstudie im Kapitel 3.4.1 und im Kapitel 3.4.2.

Gütekriterien

Die Qualität eines Tests wird traditionell an Gütekriterien wie Objektivität, Reliabilität und Validität festgemacht. In der IRT spielt die spezifische Objektivität eine zentrale Rolle. Nach dieser ist die Item- und Personenparameterschätzung unabhängig von der

Itemstichprobe und der Personenstichprobe. Der Vergleich von zwei Personen ist demnach spezifisch objektiv, wenn er unabhängig von den ausgewählten Items und den ausgewählten Personen ist. Anders ausgedrückt heißt das, die Schwierigkeitsunterschiede zwischen zwei Items lassen sich unabhängig davon feststellen, welche Fähigkeiten die zu untersuchenden Personen haben. Und Unterschiede zwischen zwei Personenparametern können unabhängig davon ermittelt werden, welches Schwierigkeitsniveau die vorgelegten Items besitzen. Dies ist eine grundlegende Voraussetzung für das adaptive Testen. Grafisch ist die spezifische Objektivität bei dichotomen Items sichtbar, wenn alle ICCs dieselbe Form aufweisen und lediglich entlang der Achse verschoben sind, auf welcher der Schwierigkeitsparameter abgebildet ist. Die spezifische Objektivität lässt sich jedoch nur den Modellen der Rasch-Familie zuschreiben (Moosbrugger, 2012). In Bezug auf die Reliabilität gibt es verschiedene Vorschläge, das Verhältnis aus wahrer und beobachteter Varianz innerhalb der IRT zu schätzen. Eine Möglichkeit ist, als wahre Varianz die als Modellparameter geschätzte Varianz der latenten Variable und als beobachtete Varianz die berechnete Varianz des Personenschätzers zu nutzen (vgl. Kapitel 3.3.1). Die Validität wird in der IRT in interne und externe Validität aufgeteilt. Die interne Validität wird dabei durch die Geltung des Testmodells abgesichert, wobei das Testmodell beschreibt, was gemessen wird. Die externe Validität wird laut Literatur nicht mehr zum Bereich der IRT gezählt (Rost, 2006).

Die Güte des Tests kann folglich danach eingestuft werden, wie gut die Qualität der Personen- und Itemparameter ist. Gerade bei individualdiagnostischen Tests stellt die Qualität der einzelnen Testergebnisse ein zentrales Gütemerkmal dar. Bei Rasch-Modellen sind beispielsweise getrennte Aussagen über die Messgenauigkeit (interne Validität) des einzelnen Testergebnisses möglich. Im Rahmen der Maximum-Likelihood ist die Messgenauigkeit z. B. über die Standardschätzfehler der Personenparameter (Schätzfehlervarianz) berechenbar. Die Varianz der Schätzwerte eines Personenparameters ist über die sogenannte Informationsfunktion berechenbar. Diese Informationsfunktion drückt aus, wie gut die in den Daten enthaltene statistische Information hinsichtlich der Schätzung eines einzelnen Modellparameters passt. Die Schätzfehlervarianz gilt für alle Parameter, die mit der Maximum-Likelihood-Methode ermittelt wurden. Der Standardschätzfehler als Qualitätsmaß eines Testergebnisses wird unter der Annahme berechnet, dass das Testmodell auf die Daten passt bzw. dass das Antwortmuster jeder

Person zum Modell passt. Zur Prüfung dieser Annahme gibt es unterschiedliche Personenfit-Indizes. Solche Indizes als Maße der internen Validität drücken aus, wie gut ein Antwortmuster zu dem zugrundeliegenden Testmodell passt. In Bezug auf die Qualität der Items werden häufig residuen-basierte oder likelihood-basierte Itemfit-Maße verwendet. Das Konzept der Itemtrennschärfe stellt ein zentrales Gütekriterium dar. Die Trennschärfe wird in der IRT als Anstieg der Itemfunktion definiert (Rost, 2004). Zusätzlich kann die klassische Trennschärfe im Sinne einer Korrelation des Items mit allen Items betrachtet werden.

3.1.2 (Computerisiertes) Adaptives Testen – Grundlagen

Es gibt grundlegende Bestimmungsstücke für einen computerisierten adaptiven Test, wie das Messmodell (vgl. Kapitel 3.1.1), ein kalibrierter Itempool (vgl. Kapitel 3.3) oder ein adaptiver Algorithmus (vgl. Kapitel 3.5). Bei der praktischen Implementierung computerisierter adaptiver Testverfahren gibt es meist zusätzliche Anforderungen an den Test zu berücksichtigen, wie z. B. die Testsicherheit oder die Balancierung von inhaltlichen Einschränkungen (Born & Frey, 2016). Diese haben zur Erweiterung des ursprünglichen Ansatzes für CAT geführt, indem z. B. die Vorgabehäufigkeit eines Items über alle Tests hinweg (*Exposure-Control*) oder die Vorgabe der Items nach inhaltlichen Vorstellungen (*Content-Balancing*) bestimmt werden kann (vgl. Kapitel 3.5.5). Diese Erweiterungen sind zum Teil schon entwickelt und zum Teil noch in Forschung. Grundlegend für jeden Ansatz ist das Verständnis des adaptiven Testens.

Adaptives Testen

„Unter *adaptivem* Testen versteht man ein spezielles Vorgehen bei der Messung individueller Ausprägungen von Personenmerkmalen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items am Antwortverhalten des untersuchten Probanden orientiert“ (Frey, 2012). Leistungsfähige Testpersonen bekommen schwierigere Aufgaben vorgelegt als Testpersonen mit mittlerem oder geringem Leistungsniveau. Das Vorgehen beim adaptiven Testen ist mit dem Vorgehen beim mündlichen Prüfen vergleichbar. Der Prüfer passt den Schwierigkeitsgrad der Fragen dem Leistungsvermögen des Prüflings an. Die vorgegebenen Items sind auf die individuelle Merkmalsausprägung des Probanden abgestimmt. Demnach wird nach jedem vorgegebenen Item die

Fähigkeit der Person θ neu berechnet und ein passendes Item (im Rasch-Modell aufgrund passender Schwierigkeit) vorgelegt. Jeder adaptive Test kann somit entsprechend der Fähigkeit der Person aus unterschiedlichen Items bestehen (Wainer & Dorans, 2000). Durch diese optimierte, der Personenfähigkeit angepasste Itemauswahl kann die Messeffizienz (Messpräzision) gesteigert bzw. die Testdauer bei gleicher Messpräzision gesenkt werden. D. h., dass bereits mit sehr wenigen Items präzise Aussagen über individuelle Merkmalsausprägungen möglich sind. So kann die Belastung für die Probanden möglichst gering gehalten werden (Asseburg, 2011). Außerdem kann durch adaptives Testen besser in den Randbereichen der Kompetenz differenziert werden als mit sequentiellen Testverfahren (Frey, 2012). So können auch sehr heterogene Gruppen, z. B. SuS beruflicher Schulen, angemessen untersucht werden. Da beim adaptiven Testen für die Itemauswahl oder die Schätzung der Personenparameter auf Grundlage der IRT komplexe mathematische Algorithmen notwendig sind, ist die Nutzung von Computern naheliegend. Nachfolgend wird deshalb adaptives Testen als computerisiertes adaptives Testen (CAT) verstanden.

Entstehungsgeschichte

Lange Zeit wurde im Bildungsbereich der Fokus nur auf papierbasierte Testungen gelegt. CAT bietet eine neue effiziente Vorgehensweise an, die vor allem durch zwei Entwicklungen begünstigt wurde. Zum einen hatte die Entstehung statistischer Grundlagen in Form der IRT seit den 1950er Jahren ihren Anteil bei der Entwicklung von computerisierten adaptiven Tests (van der Linden & Glas, 2010). Denn erst durch die Nutzung von IRT-Modellen als Messmodelle können die resultierenden Personenparameter auch bei Vorgabe unterschiedlicher Items ohne Probleme miteinander verglichen werden, wenn alle Items im Itempool die Annahmen des gewählten Modells erfüllen. Auf Basis der klassischen Testtheorie können bei adaptiver Itemvorgabe häufig keine eindeutig interpretierbaren Leistungsmaße berechnet werden (Frey, 2012). Zum anderen öffneten sich die Testentwickler in den 1980er Jahren mit der Entwicklung von leistungsstarken Computern für den Heimgebrauch für computerbasiertes Testen. Computerbasiertes Testen ermöglicht erst die effektive Nutzung von computerisierten adaptiven Tests. Weiterhin wurden in dieser Zeit im Rahmen der *Computerized Adaptive Testing version of the Armed Services vocational Aptidue Battery* grundlegende Fragen zur praktischen Anwendung adaptiven Testens untersucht. Diese Testbatterie wird beim

US-amerikanischen Militär nach wie vor zur Personalauswahl genutzt. Sie ist ein gut untersuchtes und mit ca. 400000 Probanden pro Jahr häufig verwendetes Testinstrument (Frey, 2012). Ein weiteres umfangreiches Programm ist die computerisierte adaptive Messung von Kernstandards im Bildungsbereich in den Vereinigten Staaten von Amerika (Common Core State Standards Initiative, 2010). Im deutschsprachigen Raum ist die Intelligenz-Struktur-Batterie (INSBAT) ein Beispiel für ein adaptives Testinstrument (Arendasy et al., 2009).

Vor- und Nachteile

Die Vor- und Nachteile des computerisierten adaptiven Testens ähneln denen von einfachem computerbasierten Testen in vielerlei Hinsicht (Boo & Vispoel, 1998). Die Vorteile, die sich durch computerbasiertes Testen allgemein ergeben, sind die hohe Testsicherheit, das standardisierte Testvorgehen, die probandenabhängige Testgeschwindigkeit, die schnelle und fehlerarme Testauswertung und Ergebnisrückmeldung sowie die Möglichkeit zur Verwendung innovativer Itemformate. Nachteile können ein höherer Entwicklungsaufwand im Vergleich zu papierbasierten Tests, ein hoher Aufwand bei der Bereitstellung von Computern am Testort, hohe Kosten und Probleme bei der Fairness bei computerbezogenen Personenmerkmalen sein (Frey, 2012; Linacre, 2000). CAT hat den zusätzlichen Vorteil gegenüber FIT, dass bei gleicher Messpräzision kürzere Tests bzw. bei gleicher Testlänge präzisere Tests vorliegen (Segall, 2005). Unter gewissen Voraussetzungen ist zudem eine einheitlich präzise Personenparameterschätzung über alle Fähigkeitsbereiche möglich. Dies hängt u. a. von der Beschaffenheit des Itempools und den Abbruchkriterien ab. Dazu sollten die Schwierigkeitsparameter der Items im Pool gleichverteilt sein und ein variables Abbruchkriterium genutzt werden, welches sich an der Messpräzision orientiert (Reckase, 2010). Der Hauptvorteil von CAT ist somit die erhöhte Messeffizienz als Verhältnis von Messpräzision zur Testlänge. Dabei ist die Testlänge häufig durch die Anzahl vorgelegter Items in einem Test und die Messpräzision durch den Standardfehler der geschätzten Testwerte definiert (Frey, 2012). Diese gesteigerte Messpräzision führt zu einer zuverlässigeren Messung bei Messwiederholungen und somit zur gesteigerten Reliabilität gegenüber FIT. Ein konkreter Nachteil durch CAT kann eventuell die Motivation darstellen.

Motivation

Gerade bei Testungen, wo mit einer geringeren Leistungsbereitschaft der Probanden zu rechnen ist (z. B. Erhebung von schulisch erworbenen Kompetenzen bei SuS beruflicher Schulen ohne Konsequenzen bei schlechten Testergebnissen) ist die Motivation ein wichtiger Faktor. Lange Zeit galt es als gesichert, dass adaptives Testen die Motivation zur Testbearbeitung der untersuchten Probanden steigert. Der Befund wurde damit erklärt, dass die Probanden Items vorgelegt bekommen, die auf ihr individuelles Leistungsniveau abgestimmt sind. Dadurch sollte die Vorgabe von zu leichten Items, die Langeweile auslösen können bzw. von viel zu schweren Items, die frustrieren, vermieden werden. Aktuelle Arbeiten stellen die motivationssteigernde Wirkung adaptiven Testens jedoch in Frage. Die Argumentation lautet, dass die häufig verwendete Vorgabe von Items mit mittlerer individueller Lösungswahrscheinlichkeit nicht zu einer hohen Motivation führt. Gerade leistungsfähige Personen, die in der Regel viele Items korrekt beantworten, können so im Mittel nur die Hälfte der vorgelegten Items lösen, was ungewohnt demotivierend sein kann (Asseburg, 2011). Beim Frankfurter Adaptiven Konzentrationsleistungs-Test zeigt sich, dass die Motivation bei adaptiven Testformen niedriger als bei nicht-adaptiven Testformen ausfällt (Frey, 2012). Asseburg (2011) schreibt, dass CAT aus psychologisch-motivationaler Sicht vielversprechend sein kann, insofern die Probanden zuvor über die Besonderheiten des Testalgorithmus aufgeklärt wurden. Ein praxisbezogener Hinweis ist deshalb, die Testteilnehmer ausführlich darauf hinzuweisen, dass die Items entsprechend ihrer Fähigkeit bzw. Leistung im Testverlauf vorgelegt werden und stets eine Lösungswahrscheinlichkeit von z. B. 50 % zu erwarten ist. Zudem empfiehlt Asseburg (2011) den motivationalen Effekt zu untersuchen, der entsteht, wenn die Lösungswahrscheinlichkeit von 50 % auf 70 % hochgesetzt wird.

Besonderheiten bezüglich Validität

Neben der Motivation gibt es beim adaptiven Testen auch Besonderheiten bezüglich der Validität zu beachten. Gerade bei Testentwicklung in neuen Feldern ist nicht klar, ob das entwickelte Konstrukt misst, was es angibt zu messen. Aus diesem Grund wird hier auf die Besonderheiten der Validität im Zusammenhang mit computerisiertem adaptivem Testen eingegangen. Als sehr bedeutsam scheint daher die Prüfung der *Konstruktvalidität.* Diese setzt sich zusammen aus der *konvergenten* und der *diskriminanten*

Validität. Die konvergente Validität wird durch die Korrelation zwischen verschiedenen Tests, die dasselbe Konstrukt messen, ermittelt. Konvergente Validität liegt z. B. vor, wenn die Messungen eines Konstrukts durch einen adaptiven Test mit der Messung eines Konstrukts durch einen nicht-adaptiven Test hoch miteinander korrelieren. Da bei adaptiven Tests eine höhere Messpräzision zu erwarten ist, ist anzunehmen, dass die Prüfung der konvergenten Validität zwischen unterschiedlichen adaptiven Tests stets höher ausfällt als zwischen nicht-adaptiven Tests. Die diskriminante Validität misst die Korrelation mit Tests, die ein anderes Konstrukt messen. Hier sollte die Korrelation möglichst gering ausfallen. Bei der diskriminanten Validität wird z. B. bei Leistungstests zur Messung von Maximalleistungen untersucht, ob die Maximalleistung durch Störvariablen vermindert wird. Sollte eine Korrelation zwischen den Leistungswerten und den Störvariablen vorhanden sein, deutet das darauf hin, dass die Messwerte nicht die maximale Leistung, sondern eine Mischung der Maximalleistung und der Störvariable abbilden. Störvariablen können z. B. Test- bzw. Prüfungsangst oder Lärm während der Testung sein. Testungen zur diskriminanten Validität beim Frankfurter Adaptiven Konzentrationsleistungs-Test oder bei unterschiedlichen selbstadaptierten Tests zeigen, dass die untersuchten Störvariablen bei adaptiven Testungen keinen signifikanten Einfluss hatten, hingegen bei nicht-adaptiven Testungen schon. Inwieweit diese Ergebnisse für andere adaptive Tests übertragbar sind, ist noch offen (Frey, 2012).

Neben der Konstruktvalidität spielt die *Inhaltsvalidität* eine besondere Rolle. Die Inhaltsvalidität drückt den Grad aus, in dem der Itempool insgesamt und die gewählten Items für jedes Individuum speziell die Domäne der entsprechenden Fähigkeit (z. B. Mathematikkompetenz) widerspiegeln. Im Grunde entspricht das dem Verständnis bei konventionellen Testungen. Probleme, spezifisch für CAT, können entstehen, wenn die Itemauswahlmethode nicht dem theoretischen Rahmenkonzept angepasst wird (vgl. Content-Balancing im Kapitel 3.5.5). Dies ist gesondert bei der Entwicklung adaptiver Tests zu kontrollieren (Steinberg, Thiessen & Wainer, 2000). Zudem sollte die *Kriteriumsvalidität* geprüft werden, wo die Beziehung zwischen den Ergebnissen des Tests und einem äußeren Kriterium (z. B. einem Expertenrating) quantifiziert wird. Im Wesentlichen gibt es bezüglich CAT jedoch keine Besonderheiten zu beachten (Steinberg et al., 2000).

Exkurs: Multidimensionales adaptives Testen (MAT)

Da in Testprogrammen meist mehrere Dimensionen gemessen werden sollen und es bei der Entwicklung von adaptiven Tests vor allem um die Steigerung der Messeffizienz geht, wird an dieser Stelle MAT knapp dargestellt. Multidimensionales adaptives Testen entspricht im Grundgedanken der Funktionsweise des eindimensionalen adaptiven Testens (Frey & Seitz, 2009; Segall, 1996). Jedoch werden beim multidimensionalen adaptiven Test mehrere latente Dimensionen als Ursache für das beobachtete Antwortverhalten unterstellt. So können mehrere Merkmale simultan gemessen und komplexe theoretische Annahmen mit multidimensionalen Merkmalsstrukturen direkt über das Messinstrument abgebildet werden. Als Messmodelle werden hier häufig mehrdimensionale IRT-Modelle eingesetzt (Reckase, 2009). Durch die Nutzung von multidimensionalen adaptiven Tests kann die Messeffizienz im Vergleich zu mehreren eindimensionalen adaptiven Tests gesteigert werden (Frey & Seitz, 2011). Die Messeffizienzsteigerung ist jedoch geringer, sobald suboptimale Itempools und viele Restriktionen bei der Itemauswahl vorliegen (Frey, 2012). Da es nachfolgend um unidimensionales CAT geht, wird hier lediglich auf vertiefende Literatur wie Segall (1996) verwiesen.

3.1.3 Zusammenfassung

Die IRT steht als Begriff für den Bereich der probabilistischen Testtheorie, da das Antwortverhalten der Probanden in einem Test durch ein probabilistisches Modell modelliert wird. Die logistischen Modelle lassen sich nach der Parameteranzahl unterscheiden in das Rasch-Modell (1PL), das Birnbaum-Modell (2PL) und das 3PL-Modell für dichotome und ordinale Daten. Ob das gewählte Modell im Vergleich zu konkurrierenden Modellen besser zu den Daten passt, kann beispielsweise über globale Modellgültigkeitstests (z. B. über den Likelihood-Quotient und die Chi-Quadrat Statistik) sowie über Informationskriterien analysiert werden. Da selten die Voraussetzung für inferenzstatistische Schlüsse gegeben sind, können informationstheoretische Maße als Hilfsmittel herangezogen werden. Als Gütekriterien spielen die Objektivität, die Validität sowie die Reliabilität eine Rolle. Die Güte wird innerhalb der IRT aufgrund der Personen- und Itemparameter bestimmt. Deshalb werden sogenannte Personenfit-Indizes und Itemfit-Maße verwendet.

Adaptives Testen ist ein Ansatz zur Messung individueller Ausprägungen von Personenmerkmalen, bei dem die Auswahl der vorgelegten Items am Antwortverhalten des untersuchten Probanden festgemacht wird. Der Grundgedanke besteht darin, eine optimale Passung zwischen Merkmalsausprägung und Itemschwierigkeit zu realisieren. Die Voraussetzung für CAT im hier verwendeten Sinn ist ein kalibrierter Itempool auf Grundlage der IRT und ein zuvor festgelegter adaptiver Algorithmus. CAT hat gegenüber FIT u. a. die Vorteile der gesteigerten Messeffizienz (kürzere Tests bzw. höhere Messpräzision) und der präziseren Messung in den Randbereichen der Kompetenzverteilung. Als Nachteile sind die aufwendige Erstellung und Kalibrierung des Itempools, der Mehraufwand für die Entwicklung des adaptiven Algorithmus und die zusätzliche Nutzung spezieller Computerprogramme zu erwähnen.

3.2 Testplanung

Nachdem die Grundlagen zu den Schwerpunkten IRT und CAT geklärt sind, wird in diesem Abschnitt ein theoretischer Überblick über die Testplanung gegeben. Dabei wird u. a. die Festlegung des inhaltlichen Zielkonstrukts verdeutlicht und in das Konzept der Simulationsstudien eingeführt. Als entscheidender Aspekt bei der Planung eines computerisierten adaptiven Tests wird hier auch auf Fragen zu den Themen Software und technische Umsetzung eingegangen. Nach Thompson und Weiss (2011) ist in der Testplanung die Prüfung der Durchführbarkeit des Testprogramms ein wichtiger Schritt. Es ist zu prüfen, ob mit den vorhandenen Ressourcen der gewünschte Test als adaptiver Test erstellt oder ein bestehender Test in einen adaptiven Test umgewandelt werden kann. Deshalb sollten zu Beginn die praktische und betriebswirtschaftliche Umsetzbarkeit geprüft werden. Dazu eignen sich beispielsweise folgende Fragestellungen:

- Ist ausreichend psychometrische Expertise (bezüglich IRT und CAT) vorhanden oder wird externe Unterstützung benötigt?
- Ist genügend Kapazität vorhanden, um einen umfangreichen Itempool zu erstellen?
- Bringt im konkreten Fall CAT die erwartete Verringerung der Testlänge im Vergleich zum FIT mit sich?
- Gleichen die Reduktion der Testlänge und die damit ersparten Kosten für Probandengelder die Kosten der Erstellung eines computerisierten adaptiven Tests aus?

- Sind die erhöhte Messpräzision und die gesteigerte Testsicherheit ein hinreichendes Entscheidungskriterium für CAT und gleichen sie die Mehrkosten aus?

Diese Fragen sind nicht immer ad hoc zu beantworten. Deshalb werden zuvor häufig psychometrische Studien durchgeführt. Um Antworten auf die aufgeführten Fragen zu finden, können auch Monte-Carlo Simulationen behilflich sein. Als erster Schritt der Testplanung sollte jedoch das inhaltliche Zielkonstrukt festgelegt werden.

3.2.1 Festlegung des inhaltlichen Zielkonstrukts

Dieser Schritt ist in der Testplanung hervorzuheben, da das inhaltliche Konstrukt den zu messenden Untersuchungsgegenstand und somit auch den Test an sich theoretisch bestimmt. Durch die Festlegung des inhaltlichen Konstrukts ergeben sich so immer auch Anforderungen an den konkreten Test und seinen Algorithmus. Zum einen können sich neue Aufgaben bei der Itemkonstruktion ergeben. Es müssen z. B. bei dem Ziel, mehrere Subdimensionen reliabel zu messen und auf diese Rückmeldung geben zu können, auch genügend Items in diesen Dimensionen konstruiert werden. Zum anderen hat das inhaltliche Konstrukt Einfluss auf die Itemauswahl. Wie sind die Items während der Testung zu ziehen, um bei mehreren unterschiedlichen (inhaltlichen) Anforderungen an den Test diese Anforderungen erfüllen zu können, ohne den adaptiven Algorithmus zu stark zu beeinträchtigen? An dieser Stelle kann auch die Frage aufgeworfen werden, ob ein Test, der mehrere Dimensionen misst, multidimensional oder unidimensional konstruiert werden soll. Diese Frage hat wiederum Einfluss auf die Itemkonstruktion und ggf. die Itemauswahl sowie auf das Design der Kalibrierungsstudie. Rudner (2010) gibt bezüglich der Entwicklung des *Graduate Management Admission Tests* ein Beispiel für mögliche Spezifikationen. Der Test ist inhaltlich unterteilt in drei Bereiche, welche mehrere Kategorien enthalten: Kompetenzbereich (z. B. Kategorie Problemlösen), Inhaltsbereich (z. B. Kategorie Geometrie oder Algebra) und Anwendungsbereich (z. B. Kategorie Anwendung oder Formeln). Im Zielkonstrukt wurde nun spezifiziert, dass jeder Proband innerhalb jedes Bereiches in allen Kategorien eine gewisse Anzahl an Items beantworten muss. Die Interaktion zwischen den Kategorien über die Bereiche hinweg wurde nicht berücksichtigt. Erst nachdem der Proband aus allen Kategorien genügend Items vorgelegt bekommen hat, kann der Testalgorithmus bis zum Erreichen

des Abbruchkriteriums ausschließlich nach maximaler Iteminformation (vgl. Formel (14) auf S. 61) Items auswählen.

3.2.2 Monte-Carlo Simulationen

Simulationsstudien können u. a. dabei helfen, unter den Bedingungen des inhaltlichen Zielkonstrukts die erwartete Testlänge und die damit einhergehende Präzision des Tests zu schätzen oder die benötigte Größe des Itempools bei einer gewissen Präzision des Tests vorherzusagen. Beispielsweise kann ein Test mit einem Itempool von 100 Items gegen einen Test mit einem Itempool von 200 Items simuliert werden, bevor das erste Item geschrieben wurde oder empirische Daten vorliegen. Dies ist möglich, da in Simulationsstudien die Funktionsweise eines computerisierten adaptiven Tests mit einer großen Anzahl simulierter Versuchspersonen nachgeahmt werden kann. Die Funktionsweise der adaptiven Algorithmen in der Simulation entspricht der Funktionsweise in einer empirischen Studie. Der Unterschied liegt lediglich darin, dass CAT im Feldversuch echte Probanden und deren Antworten auf die Items untersucht und bei der Simulation eine Tabelle generierter Antworten in Echtzeit vorgegeben werden. D. h., wenn der adaptive Algorithmus ein Item vorlegt, nutzt das Simulationsprogramm eine Antwort aus einem hinterlegten Datensatz (Thompson & Weiss, 2011).

Monte-Carlo Simulationen nutzen die Eigenschaften der IRT, bei der die Itemschwierigkeit b und die Personenfähigkeit θ auf der gleichen Skala abgebildet werden. Bei einem gegebenen Wert von θ kann so die genaue Wahrscheinlichkeit für eine korrekte Antwort auf ein Item i bestimmt werden. Ein Beispiel: Es wird für einen Proband mit dem Fähigkeitsschätzer $\theta = 0.0$ eine Wahrscheinlichkeit, ein Item korrekt zu beantworten, von 0.75 errechnet. Anschließend wird eine zufällige Zahl aus einer Gleichverteilung zwischen 0 und 1 gezogen. Wenn der gezogene Wert kleiner oder gleich 0.75 ist, gilt die Antwort auf das Item als korrekt, anderenfalls als inkorrekt. Als Ergebnis ist folgendes möglich: bei einer ursprünglichen Planung von 1000 Items für den Itempool mit 55 Items als Testlänge für die gewünschte Messpräzision kann die Simulationsstudie aufzeigen, dass bereits 500 Items für die gewünschte Testsicherheit und die Verteilung der Schwierigkeiten ausreichend sind und eine Kürzung der Testlänge auf 45 Items bereits die gewünschte Messpräzision im Mittel mit sich bringt. Für die Simulationsstudien sind drei Datensätze notwendig, (a) die Itemparameter für die Items im Itempool,

(b) eine Auswahl von Probandenfähigkeiten θ und (c) ein Vektor für jeden Proband mit der Angabe über korrekte und inkorrekte Antworten. Diese Datensätze können in Abhängigkeit von den zur Verfügung stehenden Daten aus empirischen Daten erzeugt oder zufällig generiert werden. Beim Generieren zufälliger Daten kann die Kritik geäußert werden, die empirische Realität nicht angemessen abgebildet zu haben. Um diese Kritik etwas zu entkräften, können z. B. die zu erwarteten Verteilungsannahmen beim Generieren der Daten mit einbezogen werden. Als abhängige Variablen bei der Monte-Carlo Simulation werden häufig die mittlere Testlänge und die Präzision bzw. der Standardfehler der Testung verwendet (Thompson & Weiss, 2011).

Für die Erzeugung der Datensätze auf Grundlage der IRT und die Simulation der Leistung des computerisierten adaptiven Tests benötigt es spezielle Software bzw. Softwarepakete für bekannte Statistiksoftware. *WinGen3* (Han, 2007) ist eine Möglichkeit, um IRT-Parameter und Antworten auf Items zu generieren. *FireStar* (Choi, 2009) oder *CATSim* (Weiss & Guyer, 2012) können genutzt werden, um CAT zu simulieren. Alternativ gibt es z. B. für das Statistikprogramm R ein kostenloses Package *catIrt (Nydick, 2014)*. Bei ausreichender psychometrischer Expertise kann solch eine Simulationssoftware auch selbst entwickelt werden. Ausgehend von den Ergebnissen der Simulationsstudien schlagen Thompson und Weiss (2011) vor, einen Plan mit Zielen und Zeit aufzustellen, um die Ziele und zukünftigen Schritte kontinuierlich daran prüfen zu können. Die Nutzung von Simulationsstudien bietet sich nicht nur zu Beginn, sondern auch während der Testentwicklung an. Beispielsweise können nach einem Pretest mit der Erhebung von empirischen Daten neue Simulationen gerechnet werden, um den adaptiven Algorithmus anzupassen. Die Nutzung von Simulationsstudien wird für den praktischen Teil an verschiedenen Stellen empfohlen (z. B. im Kapitel 4.2 bei der Entwicklung des initialen Itempools und im Kapitel 4.4 beim festlegen des CAT-Algorithmus).

3.2.3 Software und technische Umsetzung

Neben dem inhaltlichen Zielkonstrukt und der Simulation des Tests spielen bei der Testplanung eines computerisierten adaptiven Tests die Software und die technische Umsetzung eine wesentliche Rolle (Thompson & Weiss, 2011). In diesem Abschnitt werden deshalb Hauptaspekte zu Hardware- und Softwarefragen beleuchtet. Aufgrund

der Komplexität des Themas Software und technische Umsetzung sollte sich bereits zu Beginn der Testplanung mit diesem Thema beschäftigt werden. So können Erkenntnisse von Beginn an in den weiteren Testentwicklungsprozess mit einfließen. Weiterhin wird aus eigener Erfahrung empfohlen, Ressourcen für die notwendige Zusammenarbeit zwischen verschiedenen Fachbereichen von Beginn an mit einzuplanen. Beispielsweise ist eine durchgehende Interaktion zwischen inhaltlichen Experten (z. B. zur Erstellung des inhaltlichen Zielkonstrukts) und technischen Experten (z. B. zur Umsetzung des Tests in einer Software) zu gewährleisten. Zudem ist zu entscheiden, ob die Software selbst programmiert oder eine verfügbare Software auf dem Markt genutzt werden soll (Thompson & Weiss, 2011). In einer Internetrecherche wurde nach kostenloser Software zu Administration und Erstellung von computerisierten adaptiven Tests im Forschungsbereich gesucht. Als sehr umfangreiche und komfortable Software stellten sich dem Autor zwei Programme dar, welche in der nachfolgenden Tabelle kurz erläutert werden.

Tabelle 1

Eine Auswahl an Software zur Administration und Erstellung computerisierter adaptiver Tests im Überblick

Software-Name	Verfügbarkeit	Besonderheiten
Concerto: Open-Source Online Adaptive Testing Platform	freie Software	Konfiguration und Administration der Tests über den Browser; integriert in die Software R für statistische Analysen; umfangreicher mehrsprachiger Support; serverbasiert
MATE: Multidimensional Adaptive Testing Environment	in Forschung frei einsetzbar	intuitive Point- und Click-Oberfläche; lokale Konfiguration und Administration der Tests; beherrscht multidimensionales adaptives Testen; eignet sich zur einfachen Durchführung von Simulationen; kein Support

Wird eine vorhandene Software genutzt, ist von Beginn an in Betracht zu ziehen, welche technischen Restriktionen damit einhergehen und ob dadurch noch alle gewünschten Vorhaben realisierbar sind. Bei der Entwicklung einer eigenen Plattform ist genügend Zeit für die Entwicklung, Testung und Ausbesserung der Software einzuplanen (Thompson & Weiss, 2011). Bereits zu Beginn sollte der spätere Auslieferungsmodus des Tests bedacht werden. Der Itempool und/oder die Software für die Durchführung des Tests können entweder lokal auf jedem genutzten Rechner einzeln installiert oder aber zentral über ein Netzwerk (interner Server, Internet usw.) zur Verfügung gestellt werden. Beide Aspekte haben Vor- und Nachteile. Die Software lokal auf dem Computer zu speichern hat den Vorteil, dass die Rechenkraft des einzelnen PC ausschließlich für den einen Test genutzt wird. Bei einer netzwerkbasierten Lösung muss entschieden werden, ob die Software nur auf dem Server geladen oder auch ausgeführt werden soll. Sollten viele Personen gleichzeitig Testungen über das Netzwerk durchführen, braucht es ein stabiles schnelles Netzwerk und einen Server mit ausreichend Rechenleistung. Auch wenn ausschließlich der Itempool über eine Netzwerkverbindung geladen wird, können lange Ladezeiten entstehen. Diese Überlegungen können bereits zu Beginn der Testentwicklung Anhaltspunkte geben, wie aufwendig und umfangreich die Items gestaltet werden sollen (z. B. Bilder, Mediendateien usw.). Ein Vorteil der Ausführung der Tests auf dem Server kann sein, dass lokale Einstellungen eines Computers wenig Einfluss haben und die dargestellten Tests alle möglichst gleich sind. Bezüglich der Sammlung der Daten kann die Nutzung einer netzwerkbasierten Lösung ebenfalls vorteilhaft sein. So müssen nach Abschluss der Testung die Daten nicht einzeln und lokal von den Computern abgerufen und später zu einer Gesamtdatei zusammengefügt werden. Bei der Nutzung mehrerer (unterschiedlicher) Computer für die Testung, wie es beispielsweise bei einer parallelen Testung einer ganzen Klasse in einem Computerpool notwendig ist, sollte man sich bewusst sein, dass die Computer selten äquivalente Einstellungen haben. Häufig sind z. B. die benötigte Software, grafische Einstellungen, installierte Schriftarten oder die Sensibilität der Maus nicht bei allen Computern gleich. In Räumen mit vielen Computern kann zudem die Lautstärke zu einem Problem werden und die Testleistung dadurch beeinflussen. Es sind daher Computer mit passiver Kühlung oder extra lautstärkegedämpften Gehäusen zu bevorzugen. Bei Items mit Toninhalten sollten Kopfhörer verwendet werden. Ein ebenfalls nicht zu unterschätzendes Problem kann die Stromversorgung werden. Soll beispielsweise ein Klassensatz

Laptops am Testort genutzt werden, muss zuvor die Stromzufuhr sichergestellt sein. Aber auch bei der Nutzung von vorhandenen Computerpools kann es zu Stromausfall oder unbeabsichtigtem Ausschalten eines Computers kommen. Es wird deshalb empfohlen, die Testsoftware (falls möglich) so einzustellen, dass die Ergebnisse nach jeder Antwort zwischengespeichert werden (Green, 2000).

Die Frage nach den Eingabegeräten für die Antworten kann ebenfalls Einfluss auf die Testentwicklung haben. Eventuell sind geplante Itemformate anzupassen. Ein häufig genutztes Standard-Eingabegerät, gerade bei offenen Textantworten, ist die Tastatur. Bei ausschließlicher Nutzung von Single-Choice bzw. Multiple-Choice-Items ist eine Computermaus meist ausreichend und kann bei einigen Testsystemen die Sicherheit erhöhen. Beispielsweise wird eine Internetrecherche ohne angeschlossene Tastatur erschwert. Aus technischer Sicht sind auch Spracheingaben und Eingaben über Touchscreen bzw. über einen elektronischen Stift denkbar. Systeme, welche beispielsweise die Augen-Handkoordination über Videokameras erfassen sind ebenfalls möglich (Strain-Seymour, Walter & Robert, 2009). Bei der Nutzung des jeweiligen Eingabegerätes sollte auch darauf geachtet werden, dass die Testsituation dadurch möglichst nicht beeinflusst wird (z. B. für Linkshänder ein problemloses Umstellen der Maus ermöglichen).

Bei der Wahl des Bildschirms gibt es ebenfalls einige Dinge zu beachten. Es sollten möglichst keine Unterschiede zwischen unterschiedlichen Probanden geben. Änderungen in der Displaygröße oder der Qualität können die Lesegeschwindigkeit und das Antwortverhalten beeinflussen. Bereits bei der Erstellung der Items ist darauf zu achten, dass die Inhalte an den gewählten Displays gut lesbar sind. Die Buttons, Eingabefenster, Schriftgrößen usw. sollten immer proportional je nach Displaygröße einheitlich sein, so dass diese sich durch unterschiedliche Displays nicht verschieben. Bei einem Item mit umfangreichem Text oder vielen Bildern kann es notwendig sein, das Item entweder auf mehrere Seiten aufzuteilen oder aber das Item so zu konstruieren, dass es über die Scroll-Funktion der Maus oder Tastatur lesbar ist. Matte Displays mit einer hohen Auflösung und hoher Helligkeit mindern Spiegelungen auf dem Bildschirm durch Licht und Sonne. Abgedunkelte Räume sind jedoch zu bevorzugen. So wird der Kontrast des Displays wenig beeinflusst (Green, 2000).

Gerade bezüglich der Interaktion Computer Mensch sind beim computerisierten adaptiven Testen besondere Dinge zu beachten (vgl. Kapitel 3.1.2). Wise und Kingsbury (2000) berichten von spezifischen Aspekten, die CAT betreffen und welche die Testleistung eines Probanden beeinflussen können. Sie diskutieren u. a. den Aspekt des *Item-Review*, wodurch während der Testung im Test zurückgegangen werden kann. So können Items noch einmal beantwortet bzw. eine bereits gegebene Antwort geändert werden. Studien zeigen, dass Probanden strategisch die Möglichkeit des Item-Review nutzen können, um ihren Punktwert und somit die geschätzte Leistung in der Testung zu erhöhen. Ein Nachteil des Item-Review ist, dass zusätzliche Testzeit eingeplant werden sollte. Denn das Zurückgehen und Ändern der Antworten verbraucht zusätzlich Zeit ohne mehr Testgenauigkeit zu erbringen. Der Standardfehler könnte somit für dieselbe Testzeit wesentlich höher ausfallen als im Vergleich zu einem Testverfahren ohne Item-Review. Ein Verhindern des Item-Review kann hingegen zu einem Gefühl von Kontrollverlust beim Probanden und somit zu Angst- und Stresssituationen führen, die sich negativ auf den Testablauf auswirken können. Diese Aspekte sollten bei der Wahl für oder gegen das Item-Review-Verfahren berücksichtigt werden.

Aber nicht nur die Frage nach dem Zurückgehen im Test, sondern auch nach dem Weitergehen zum nächsten Item ist zu klären. Es ist sinnvoll, eine Zeitverzögerung einzubauen, wenn das Item-Review verboten wird. D. h., nach dem Erscheinen des Items auf dem Bildschirm kann nicht direkt zum nächsten Item weitergegangen werden. Erst nach Ablauf einer gewissen Zeit (z. B. nach vier Sekunden) wird der Button sichtbar und kann gedrückt werden. So kann ein versehentliches Weiterklicken bzw. ein einfaches Durchklicken vermieden werden. Alternativ dazu besteht die Möglichkeit, ein Weitergehen zum nächsten Item erst zu ermöglichen, nachdem eine Antwort gegeben wurde. Falls ein Speed-Test konstruiert werden soll, ist die Zeit, bis man Weiterklicken darf, zu berücksichtigen (Green, 2000). Für den gesamten Test empfehlen Wise und Kingsbury (2000) auf ein Zeitlimit zu verzichten. Da durch CAT bereits Zeit eingespart wird und die Geschwindigkeiten der Probanden sehr unterschiedlich sind, sollten die Probanden die Zeit bekommen, die sie brauchen, um z. B. einen gewissen Standardfehler bzw. eine gewisse Anzahl an Items zu erreichen. Dies verhindert die Beeinflussung der Testperformanz durch Angst und Stress und verbessert die Testvalidität (insofern Zeitdruck nicht im zu messenden Konstrukt vorgesehen ist).

Als letzter Punkt wird empfohlen, die Probanden über die Besonderheiten beim Ablauf und während des adaptiven Tests zu informieren. Beispielsweise kann der Proband vorab darüber informiert werden, dass die Lösungswahrscheinlichkeit der Items an die Fähigkeit der Personen im Test angepasst ist und der Test deshalb häufig als schwer empfunden wird. Das kann sich positiv auf die Motivation auswirken (Asseburg, 2011). In der Instruktion sollte auf technische Aspekte wie Länge des Tests, Zeitlimits, Item-Review, Weitergehen im Test nach erfolgter Antwort, Scrolling, Abbruchkriterien usw. hingewiesen werden. Auch die Nutzung von Checkboxen, Radiobuttons, offenen Textfeldern sollte erklärt werden. Da es sich um ein interaktives System handelt, sollte jede Antwort eines Probanden außerdem als Änderung im Display z. B. über Hinweisboxen oder Markierungen sichtbar werden (Green, 2000).

3.2.4 Zusammenfassung

Neben allgemeinen Herausforderungen, die es beim computerisierten Testen zu beachten gibt, ist gerade beim adaptiven Testen die Interaktion zwischen Mensch und Computer zu berücksichtigen. Im Abschnitt Testplanung wurde ein theoretischer Überblick über die Schritte gegeben, die von Beginn an in die Testentwicklung einfließen sollten. Der Prozess der Testplanung kann durch die Nutzung von Monte-Carlo Simulationsstudien erheblich vereinfacht werden. Wichtige Fragen zur Durchführbarkeit und zur Aufstellung der Ziele können mithilfe von Simulationsstudien beantwortet werden. Es wurden deshalb Möglichkeiten aufgezeigt, um selbständig Simulationsstudien durchführen zu können. Auf die Bedeutung des inhaltlichen Zielkonstrukts in der Testplanung wurde hingewiesen und dabei auf Besonderheiten für CAT eingegangen. Als ein Hauptpunkt bei computerisierten adaptiven Tests wurden Softwarefragen und Fragen zur technischen Umsetzung behandelt. Als eine Software, die sowohl die Simulation von adaptiven Tests beherrscht als auch als Testplattform genutzt werden kann, bietet sich die Software MATE an. MATE wurde als Administrationssoftware im empirischen Teil dieser Arbeit verwendet. Bereits bei der Testplanung sollte berücksichtigt werden, dass ein computerisierter adaptiver Test der Wartung und der Pflege bedarf, wenn er über einen längeren Zeitraum genutzt werden soll (vgl. Kapitel 3.6.3). Dies betrifft auch die Verwaltung der Testsoftware. Sollten bei der Wartung Änderungen am Itempool oder am adaptiven Algorithmus erfolgen, bedeutet dies auch stets Änderungen in der verwendeten Software vorzunehmen. Der Testentwickler sollte deshalb sicherstellen, dass

entweder ein Support für die Software auch nach der Testentwicklung besteht oder er selbst die Fertigkeiten und Rechte besitzt, die Änderungen selbstständig vorzunehmen.

3.3 Entwicklung des initialen Itempools

Nach der Testplanung kann die Entwicklung des initialen Itempools für den adaptiven Test beginnen. Die Qualität des Itempools ist entscheidend für das Funktionieren des adaptiven Algorithmus. Auch ein hervorragender adaptiver Algorithmus kann eine zu geringe Anzahl an Items oder schlechte Qualität von Items nicht ausgleichen (Flaugher, 2000). In diesem Abschnitt werden deshalb Anforderungen besprochen, die CAT an den Itempool stellt. Zudem werden Aspekte der Itementwicklung speziell für computerisierte Items bzw. Items in computerbasierten Testungen beleuchtet. Bei der Entwicklung des Itempools wird empfohlen, die Items in einer elektronischen Itemdatenbank zu sammeln. Vale (2006) zeigt einen konzeptionellen Ansatz, um wichtige Aspekte bei der Auswahl und dem Design einer elektronischen Itemdatenbank zu berücksichtigen. Die elektronische Itemdatenbank enthält neben der organisierten Sammlung der einzelnen Items aus dem Itempool weitere Informationen, z. B. zu Subdimensionen (Inhaltsbereichen), kognitiven Anforderungen, Antworttypen, Itemparametern, Angaben zum Scoring oder anderen wichtigen Kriterien. Zudem sollte jedes Item eine einzigartige ID besitzen (eindeutige Identifikation). Weiterhin können Beziehungen zwischen den Items oder zwischen Kriterien der Items in der Datenbank, der Name des Itementwicklers, die Nutzungshistorie der Items, Quellenangaben oder statistische Kennwerte zur Skala hinterlegt werden. Es gibt unterschiedliche Möglichkeiten, die Itemdatenbank mit Items zu füllen. Häufig wird ein existierender Itempool als Ausgangspunkt genutzt und dieser für CAT angepasst. Aber auch das Sammeln von Items aus verschiedenen Testungen, die inhaltlich das gleiche Zielkonstrukt messen, ist möglich (Reckase, 2010). Eine weitere Möglichkeit ist es, Items komplett neu zu entwickeln. Flaugher (2000) stellt einen allgemeinen Plan zur Entwicklung eines Itempools vor:

- Erstelle eine suffiziente Anzahl an Items in jeder inhaltlich zu untersuchenden Kategorie des inhaltlichen Zielkonstrukts, basierend auf den zu erfüllenden Testspezifikationen (z. B. angestrebte Verteilung der Schwierigkeiten).
- Überprüfe die Items auf Qualität.

- Führe einen Pretest für die neu geschriebenen Items durch.
- Entferne bzw. überarbeite Items, die aufgrund der Ergebnisse des Pretests und statistischer Itemanalysen (konventionell und auf Grundlage der IRT) als unpassend erscheinen.
- Sollte der computerbasierte Test zuvor in anderer (z. B. papierbasierter) Form durchgeführt worden sein, vergleiche die Verteilung des resultierenden Itempools mit der Verteilung des Itempools der vorherigen Testform und evaluiere mittels Simulationsstudien die Funktionsweise des Content-Balancing in den unterschiedlich möglichen Fähigkeitsbereichen der Probanden.
- Wandel die Items in eine computerisierte Form um.

Dies ist ein sehr allgemeines Vorgehen, welches nicht zwangsläufig auf jede Studie zutreffen muss und welches wenig über die Anforderungen des Itempools speziell beim adaptiven Testen aussagt.

3.3.1 Anforderungen des Itempools

Was sind also die Anforderungen an den Itempool und die Items in einem Itempool? Wie groß muss beispielsweise der Itempool für eine geplante Studie sein? Wie müssen die Items im Itempool verteilt sein? Die Größe und die Verteilung eines Itempools hängen von dem Design des adaptiven Tests und der Verteilung der Leistungsparameter in der Zielpopulation ab. Reckase (2010) schlägt eine Prozedur vor, um die Anforderungen an einen Itempools für CAT für die konkrete Studie zu ermitteln, um den adaptiven Algorithmus optimal zu unterstützen. Die vorgeschlagene Prozedur funktioniert auch in Verbindung mit Content-Balancing, Exposure-Control und unterschiedlichen Itemauswahlmethoden für einparametrige Modelle in der Praxis gut. Anforderungen an einen Itempool, um den adaptiven Algorithmus optimal zu unterstützen und somit die Messpräzision möglichst hoch zu halten sind nach Urry (1977) die Itemdiskrimination, die Verteilung der Itemschwierigkeit, der Rateparameter und die Anzahl der Items im Pool. Demnach soll für ein Item i die Itemdiskrimination a_i höher als 0.8 sein, der Itemschwierigkeitsparameter b_i eine Breite der Verteilung von mindestens -2.0 bis $+2.0$ haben, der Rateparameter c_i kleiner als $.3$ sein und mindestens 100 Items im Itempool enthalten sein. Diskrimination, Schwierigkeit und Ratewahrscheinlichkeit der Items lassen sich jedoch nur bedingt bei der Itemerstellung beeinflussen. Diese Parame-

ter können mit Sicherheit erst nach der Kalibrierungsstudie festgelegt werden. Die endgültige Itempoolgröße ist ebenfalls nur schwer vorherzusagen, da nach der Kalibrierung der Items und dem damit einhergehenden Pretest häufig Items aus unterschiedlichen Gründen aus dem Itempool entfernt werden müssen. Wise und Kingsbury (2000) berichten, dass die Größe von 100 Items nach wie vor eine gute Poolgröße für CAT ist, dass jedoch aktuelle computerisierte adaptive Tests auf Itempools von mehr als 1000 (teilweise mehr als 2000) Items zurückgreifen. Sie sehen dafür drei Gründe:

1) Die konventionellen Testungen sind in den letzten Jahrzehnten deutlich besser geworden und kommen teilweise an die Präzision eines zweistufigen adaptiven Tests heran.

2) Es gab eine Entwicklung verschiedenster Möglichkeiten, Restriktionen an den Test zu stellen (z. B. inhaltliche Restriktionen aufgrund des inhaltlichen Zielkonstrukts der Testung).

3) Testungen, bei denen die Testsicherheit wichtig ist (z. B. Prüfungen) benötigen große Itempools, um die Häufigkeit des Auftauchens eines Items über die Zeit hinweg zu kontrollieren. Bei einem kleinen Itempool ist es für die Probanden wesentlich einfacher, sich alle Items zu merken und an andere Probanden weiterzugeben als bei einem sehr großen Itempool.

Deshalb sollten bei der Erstellung des Itempools für einen adaptiven Test nicht nur der allgemeine Itemauswahlmechanismus und die damit einhergehende mögliche Reliabilität als Referenzkriterium für die nötige Anzahl an Items dienen. Je nach Zielvorstellungen sind z. B. auch Aspekte der Testsicherheit oder der Inhaltskontrolle im eigenen Testprogramm zu beachten und haben somit Einfluss auf die nötige Anzahl an Items im Itempool. An dieser Stelle bietet es sich an, Simulationsstudien zu verwenden. Weiterhin ist zu berücksichtigen, dass nach dem Pretest und der Kalibrierung der Items in der Regel noch Items aus dem Itempool entfernt werden. Die Anzahl der zu entfernenden Items unterscheidet sich je nach Studie. Die Komplexität des zu messenden inhaltlichen Zielkonstrukts, die Verwendung und Anpassung eines vorhandenen Itempools gegenüber der Entwicklung eines neuen Itempools, die Festlegung der Ausscheidungskriterien (z. B. Differential Item Functioning, Signifikanzniveaus usw.)

und viele andere Faktoren bestimmen darüber, wie viele Items nach dem Pretest aus dem Itempool entfernt werden. Bei der Entwicklung des Itempools spielt auch die Dimensionalität der Itemantworten eine Rolle. Die Dimensionalität sollte bei der Itementwicklung bereits mitgedacht und nach der Kalibrierungsstudie geprüft werden (Wise & Kingsbury, 2000). Für CAT sollten bestenfalls unidimensionale Itempools für die einzelnen zu messenden Dimensionen erstellt werden, um den adaptiven Algorithmus zu unterstützen (Flaugher, 2000). D. h., die Itemantwort auf ein Item sollte zweifelsfrei einer Dimension zugeordnet werden können und sich auch in dem entsprechenden Itempool wiederfinden. Sollten spätere Analysen darauf hindeuten, dass das zu messende Konstrukt multidimensional ist und die Itemantworten Rückschlüsse auf mehreren Dimensionen zulassen, muss auch darauf geachtet werden, später ein passendes multidimensionales Item Response Modell zu wählen (Wise & Kingsbury, 2000). Nachdem die Anforderungen des Itempools definiert sind, müssen die einzelnen Items in eine computerisierte Form gebracht werden.

3.3.2 Entwicklung von Items für CAT

An dieser Stelle wird es keine ausführliche Anleitung zur Konstruktion von Testitems geben, da mit dieser Arbeit eine zeit- und ressourcensparende praktische Anleitung zur Testentwicklung entwickelt wurde. Die Nutzung bestehender Items (Itemrecycling; vgl. Kapitel 4.2) ist deshalb ein wesentlicher Bestandteil. Dennoch wird es Bereiche geben, in denen man nicht umhinkommt, Items neu zu entwickeln. Aus diesem Grund wird auf das Handbuch zur Testentwicklung von Haladyna (2004) und das Werk von Osterlind (1998) zur Itemkonstruktion hingewiesen. Nachfolgend werden auf die wichtigsten Aspekte von (innovativen) Items und auf die Besonderheiten von Items in computerbasierten Testungen sowie auf Möglichkeiten der Nutzung innovativer Items eingegangen.

> A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as an knowledge, ability, predisposition, or trait) may be inferred (Osterlind, 1998, S. 19).

Vale (2006) schreibt, dass ein Item mehr ist, als eine bloße Fragestellung. Ein Test besteht selten nur aus einer Sammlung von Fragen. Häufig sind neben den Fragestellungen auch komplexe Probleme zu lösen oder aufgestellte Behauptungen zu bewerten. Ein gutes Item sollte deshalb als Grundlage stets einen Stimulus enthalten. Als Stimulus wird hier eine Frage, ein Statement, eine Abbildung, eine Tabelle oder eine andere Form gemeint, in der Informationen gegeben, Probleme aufgezeigt oder das Denken angeregt werden können. Bei computerbasierten Testungen und somit auch bei computerisierten adaptiven Tests sind Videos, bewegte Grafiken oder Sound-Elemente als zusätzliche Elemente einfach einsetzbar. Die Nutzung von Medien (Grafik, Video, Animation und Sound) kann beispielsweise dazu dienen, den Leseaufwand während der Testung zu verringern, was zu einer kürzeren Testzeit führen kann (Vale, 2006). Prinzipiell ermöglicht der Einsatz des Computers die Nutzung von innovativen Items.

Strain-Seymour et al. (2009) stellen einen Ansatz vor, um innovative Items kostengünstig und zeitsparend zu erstellen. Entscheidend für den vorgestellten Ansatz ist, dass er eine effiziente Strategie zur Itemerstellung darstellt. Der Ansatz zielt auf geringe Kosten für die Itementwicklung und einen hohen Grad an Bedienerfreundlichkeit der Items bei hoher Itemqualität ab. Der Kern des Ansatzes besteht aus der Verwendung von Elementvorlagen (Item-Templates). Item-Templates sind wiederverwendbare Modelle oder Muster, mit denen schnell individuelle Vorlagen von Items erstellt werden können. Die Flexibilität der Templates wird dadurch gewährleistet, dass die Item-Elemente stets wiederverwendet werden können. So werden Programmierkosten reduziert, Zeit gespart und eine unabhängige Arbeit der Inhaltsexperten von der technischen Umsetzung gewährleistet. Innovative Items bieten einige Vorteile. Mit innovativen Items können (a) ein breiteres Spektrum an Fähigkeiten als mit einfachen Items gemessen, (b) die Authentizität der Testsituation gesteigert, (c) komplexe und dynamische Informationen präsentiert, (d) die Lesebelastung verringert, (e) die Bereitschaft der Probanden gesteigert, (f) die Ratewahrscheinlichkeit und die Anforderungen an das Arbeitsgedächtnis gesenkt und so die Validität der Messung gesteigert sowie (g) die Prozesskompetenz gemessen werden (Strain-Seymour et al., 2009). Die Nutzung von Templates bietet beispielsweise die Möglichkeit, konkrete Aufgaben durch den Computer während der Testung erstellen zu lassen. Bei einem einfachen Mathetest können so beispielsweise die Zahlen in einer Aufgabe im Verlauf der Testung zufällig eingefügt und die korrekten

Ergebnisse durch den Computer ermittelt werden (Parshall, Harmes, Davey & Pashley, 2010).

Bei der Nutzung innovativer Items lassen sich auch die Interaktionsmöglichkeiten zwischen Items und Probanden erweitern. Beispielsweise ist eine kontinuierliche Interaktion, wie in einem Computerspiel, bei einer Testung denkbar. Oder aber die Interaktion kann anstatt über die Eingabe per Maus und Tastatur über eine Kamera erfolgen. Auf diese Weise können ganz neue Personengruppen und Fähigkeiten untersucht werden. Prinzipiell ist bei der Entwicklung von Items das Thema Interaktion zentral. Hier stellt sich stets die Frage, ob das Item kognitiv noch das misst, was gemessen werden soll. Zudem ist bei dem parallelen Einsatz von papierbasierten Testungen zu ermitteln, ob Items nicht aufgrund der unterschiedlichen Interaktionsformen (Stift und Tastatur bzw. Maus) unterschiedlich funktionieren und ggf. Personengruppen bevorteilen (Strain-Seymour et al., 2009). Eine Herausforderung stellt auch die zunehmende Komplexität bei der Zunahme möglicher Elemente innovativer Items dar. Hieraus ergibt sich die Gefahr, dass überfrachtete Schnittstellen zwischen Computer und Proband oder nicht ergonomisch programmierte Software dazu führen, die Items unnötig komplex zu gestalten. Bei der Nutzung realistischer Testungen (z. B. Flugzeugsimulationen, Erste-Hilfe-Simulationen usw.) müssen nicht alle Möglichkeiten ausgeschöpft werden, um bestimmte Ergebnisse zu messen (Parshall et al., 2010). Die optische und akustische Simulation einer Person mit schweren Schmerzen bei einer allgemeinen Wissenstestung im Bereich Erste Hilfe ist zwar möglich, aber meist unnötig. Dadurch steigen lediglich die Komplexität der Items und die Kosten bei der Entwicklung an.

Ein weiterer wesentlicher Punkt bei der Erstellung von Items ist die Wahl des Antwortformates. Die Antwort auf ein Item muss nicht zwangsläufig dichotom (z. B. korrekt und nicht korrekt) ausfallen. Es sind Itemformate möglich, die je nach Anzahl richtiger Antworten mit keinem Punkt, einem Punkt oder mehreren Punkten bewertet werden können. In Anlehnung an die Unterteilung der Itemantwortformate von Parshall et al. (2010) werden im empirischen Teil dieser Arbeit folgende Antwortformate verwendet: (a) geschlossene Antwortmöglichkeiten (Selected Response Items) und (b) offene Antwortmöglichkeiten (Constructed Response Items). Die geschlossenen Antwortmöglichkeiten werden unterschieden in Single-Choice-Items, Multiple-Choice-Items und komplexe Multiple-Choice-Items. Die offenen Antwortmöglichkeiten werden differen-

ziert zwischen einfachen offenen Formaten und komplexen offenen Formaten. Konkret werden die Formate wie folgt definiert:

- *Single-Choice Antwortformat*: Dieses Format ist auch als einfaches Multiple-Choice Format bekannt. Meist gibt es einen Stimulus und zwei bis fünf unterschiedliche Antwortmöglichkeiten. Nur eine Antwortmöglichkeit ist korrekt, die restlichen Antwortmöglichkeiten (sogenannte Distraktoren) sind falsch. Es kann genau eine Antwort ausgewählt werden.
- *Multiple-Choice Antwortformat:* Items in diesem Format werden auch als Multiple Response Items bezeichnet. Es können mehrere Antworten ausgewählt werden und richtig sein.
- *Komplexes Multiple-Choice Antwortformat:* Zu jeder Antwortmöglichkeit gibt es mehrere Möglichkeiten darauf zu reagieren (z. B. richtig, teilweise richtig, falsch). Dieses Vorgehen wird auch als eindeutige Antwortauswahl bezeichnet, da zu jeder Antwortmöglichkeit bewusst Stellung genommen werden muss. Hier können ebenfalls mehrere Antworten richtig sein.
- *Einfaches offenes Antwortformat:* Es kann als Antwort auf eine Frage ein begrenzter freier Text eingegeben werden. Alle möglichen korrekten Antworten müssen in einer Datenbank hinterlegt sein oder müssen mittels Syntax durch den Computer ermittelt werden können, um als richtig interpretiert zu werden. Die Komplexität der richtigen offenen Antwort ist dadurch sehr beschränkt.
- *Komplexes offenes Antwortformat:* Hier ist eine komplexe freie Antwort möglich (z. B. ausführliche Begründungen über mehrere Sätze hinweg). Die Antworten können meist erst nachträglich bewertet werden, da automatisierte umfangreiche Bewertungsalgorithmen fehlen oder zu aufwendig sind.

Bei der Wahl des Antwortformates kann sich der Entwickler folgende Fragen stellen: Benötigt der Proband (zusätzliche) Computerkenntnisse, um das Item zu lesen, mit ihm zu interagieren oder es zu beantworten? Ist der Antwortbereich einfach verständlich, um alle Probanden zu befähigen, effizient eine Antwort zu geben? Ist die Anleitung bzw. der Hinweis zur Beantwortung des Items unter Berücksichtigung der innovativen Iteminhalte klar und ausführlich genug? Wichtig bei allen verwendeten Iteminhalten ist, dass die Bewertung einer Itemantwort bei einem adaptiven Test bei dem Großteil der Items

automatisch erfolgen sollte, damit die Information der Bewertung in die Itemauswahl für das nächste Item einfließen kann (Parshall et al., 2010). Außerdem sollte nach Möglichkeit jeder Stimulus genau einem Item zugeordnet werden. Gilt ein Stimulus für mehrere Items, wird dies häufig als *Testlet* bzw. *Itemcluster* bezeichnet. Dies ist eine häufig genutzte Methode, um Entwicklungs- und Testzeit zu sparen. Die Nutzung von Testlets bietet sich beispielsweise in der Domäne Lesen beim FIT an, wo häufig lange Text-Stimuli notwendig sind, aber Testzeit gespart werden soll. Die Verwendung von Testlets kann unter Umständen aber dazu führen, dass die Annahme der lokalen stochastischen Unabhängigkeit in der IRT nicht erfüllt wird (Flaugher, 2000). Beim adaptiven Testen würden zudem bei stark abweichenden Itemschwierigkeiten innerhalb eines Testlets immer auch Items mit wenig passender Schwierigkeit vorgelegt werden, was zu Einbußen in der Messeffizienz führt. Als praktischer Hinweis wird in Bezug auf den Stimulus deshalb empfohlen, möglichst keine Testlets zu verwenden oder darauf zu achten, dass die Items innerhalb der Testlets annähernd gleiche Schwierigkeiten besitzen.

Eine zusätzlich hier erwähnte Möglichkeit bei der Nutzung computerbasierter Items ist die Verwendung von Ergebnisprotokollen, den sogenannten *Log-Daten*. In den Log-Daten werden Informationen gespeichert, die Auskunft darüber geben, wie die Items im Test bearbeitet wurden. Dort können z. B. die Häufigkeit des Anhörens einer Tonspur, der Zeitpunkt einer Pause im Video, die Bearbeitungszeit für ein Item oder die Anzahl an benötigten Mausklicks zum Markieren einer Grafik enthalten sein. Alle diese Informationen können genutzt werden, um die Interaktion zwischen Proband und Item zu untersuchen.

3.3.3 Zusammenfassung

In diesem Abschnitt wurde auf die Erstellung eines initialen Itempools für CAT eingegangen. Dazu wurde ein allgemeines Vorgehen für die Erstellung eines Itempools vorgestellt und wichtige Anforderungen an einen Itempool für computerisiertes adaptives Testen wie z. B. die Anzahl der Items oder die Dimensionalität der Items besprochen. Zudem wurde der Aufbau eines Items festgelegt. Weiterhin wurden verschiedene Möglichkeiten innovativer Items gezeigt. Zur Erstellung von Items wurde der Template-basierte Ansatz dargestellt. Für eine einheitliche Darstellung wurden die

Itemantwortformate Single-Choice, Multiple-Choice, komplexes Multiple-Choice, einfache offene Antwortmöglichkeiten und komplexe offene Antwortmöglichkeiten definiert.

3.4 Pretest und Kalibrierung des Itempools

Neben den Items im Itempool benötigt der adaptive Algorithmus feste Itemparameter zur Itemauswahl und zur Merkmalsschätzung. Diese werden in der Regel vorab durch eine Kalibrierungsstudie empirisch ermittelt. Kalibrierung bezeichnet hier somit die Festlegung der Itemparameter. In einer Kalibrierungsstudie für einen computerisierten adaptiven Test werden meist viele Items kalibriert, so dass diese nicht alle einer Person vorgelegt werden können. Aus diesem Grund wird oft ein Design verwendet, welches die Anordnung der Items festlegt (Testheftdesign). Die Kalibrierungsstudie dient neben der Kalibrierung der Items häufig auch als Pretest. Beim Pretest wird u. a. die Qualität des Itempools geprüft. Bei dieser Prüfung werden häufig Items aus dem Itempool entfernt, weil sie z. B. nicht zum gewählten Modell passen oder aufgrund von Differential Item Functioning (DIF) nicht geeignet sind. In diesem Kapitel wird es eine Einführung in die Themen Testheftdesign und Kalibrierungsstudie geben. Dabei werden Fragen bezüglich der Itemparameterschätzung, der Passung der Items zum gewählten Modell (Modellfit) und des DIF beantwortet. Zudem wird in das Themenfeld der Itempositionseffekte eingeführt. Itempositionseffekte können Einfluss auf die Itemparameterschätzung haben. Um damit angemessen umgehen zu können, müssen bereits bei der Testheftplanung Überlegungen angestellt werden.

3.4.1 Testheftdesign und Kalibrierungsstudie

In der Kalibrierungsstudie geht es neben dem Pretest zur Prüfung der Itemgüte (vgl. Kapitel 3.4.2) um die Schätzung der Itemparameter innerhalb des gewählten IRT-Modells. Die festen Itemparameter werden im adaptiven Algorithmus u. a. für die Itemauswahl benötigt (Eggen & Verhelst, 2011). Bei der Planung der Kalibrierungsstudie sollten vorab relevante Einflussfaktoren auf die Schätzung der Itemparameter berücksichtigt werden. Dabei ist z. B. zu unterscheiden, ob eine neue Metrik eingeführt wird oder ob auf einer bestehenden Metrik berichtet werden soll. Wenn ein Itempool komplett neu erstellt wird, ist davon auszugehen, dass auch auf einer neuen Metrik

berichtet wird. Wenn bereits kalibrierte Items im Itempool vorhanden sind und daraus bereits eine Metrik erzeugt wurde, können die Itemparameter auf Grundlage dieser Metrik ermitteln werden. Glas (2010) unterscheidet bei der Kalibrierung zwei Stufen. Die erste Stufe beschreibt die *pretesting stage*, in der ein Teil der Items einem Teil von Probanden vorgelegt wird, um die Parameter grundlegend für die Nutzung im adaptiven Test zu erheben. Die zweite Stufe, die *online stage*, beschreibt das Vorgehen, bei dem bereits Items mit Itemparametern im Itempool vorliegen und live während des adaptiven Testens neue Items hinzugefügt werden sollen. So können die Informationen aufgrund der geschätzten Personenfähigkeit genutzt werden, um weitere Itemparameter zu schätzen. In diesem Kapitel wird lediglich das Vorgehen der ersten Stufe beschrieben. Das Prüfen von Itemparametern über die Zeit oder das Hinzufügen von neuen Items zu einem Itempool wird später näher erläutert (vgl. Kapitel 3.6.3). Ein weiterer Einflussfaktor auf die Schätzung der Itemparameter ist die vorhandene Anzahl an Antworten pro Item. Dies hat Einfluss auf die Genauigkeit der Schätzung und beeinflusst die Planung des Testheftdesigns (Thompson & Weiss, 2011). Weitere Faktoren, welche die Schätzung der Itemparameter im Rahmen der IRT beeinflussen können sind Stichprobengröße (Anzahl an Probanden), Testlänge (Anzahl an Items), Verteilung des Fähigkeitsparameters (z. B. normalverteilt) oder die verwendete Methode zur Itemparameterschätzung (z. B. Maximum Likelihood-Schätzung). Ein Überblick über die genannten Faktoren und die Auswirkungen auf Gütefaktoren der Items finden sich bei Yoes (1995).

Da der Itempool für einen adaptiven Test sehr groß sein kann, erhält oftmals nicht jeder Proband jedes Item in der Kalibrierungsstudie. Deshalb empfiehlt es sich, ein durchdachtes Testheftdesign zu nutzen, um jeden Probanden eine festgelegte Auswahl an Items vorgeben zu können. Das Testheftdesign bezeichnet hier die konkrete Anordnung der Items zu jedem möglichem Testheft. Als Testheft wird in dieser Arbeit eine vor dem Test festgelegte Anordnung bzw. Abfolge von Items definiert, welche in der Kalibrierungsstudie mittels Computer in Form eines FIT vorgegeben wurde. Dabei ist anzumerken, dass der Begriff Testheft ursprünglich aus dem Bereich der papierbasierten Testung kommt. Im Kontext der computerbasierten Testung ist auch der Begriff Testzusammenstellung möglich, welcher hier äquivalent zum Begriff des Testhefts zu sehen ist (Frey et al., im Druck). In einem Testheft können die Items randomisiert oder nach einer vordefinierten Anordnung zugeordnet werden. Zeit- und Motivationsgründe sind

Ursachen dafür, unvollständige Designs zu nutzen (Eggen & Verhelst, 2011). Frey et al. (2009) empfehlen ein balanciertes unvollständiges Design als Testheftdesign zu verwenden, um z. B. Itempositionseffekte in Testungen statistisch kontrollieren zu können. Die statistische Kontrolle des Positionseffektes erfolgt durch die Mittelung der ungewollten Variabilität der Parameterschätzung über die Positionen hinweg. Eine Möglichkeit, Positionseffekte und weitere Faktoren im Testheftdesign zu berücksichtigen ist das Youden-Square-Design (YSD). Ein YSD ist ein balanciertes unvollständiges Blockdesign (BIBD) für t Treatments (hier Einzelitems) in b Blocks (hier Testhefte). Dabei taucht jedes Item t höchstens einmal in einem Testheft b auf, jedes Item erscheint genau r mal über alle Testhefte hinweg. Jedes Testheft hat eine identische Länge k, wobei $r = k$ ist. Jedes Paar von Items taucht in den Testheften maximal mit einer Frequenz von λ auf. Das hat zur Folge, dass jedes Item in jedem Testheft und auf den Positionen innerhalb des Testheftes gleich häufig erscheint (Frey et al., 2009). Die Verwendung des YSD ermöglicht es somit, für jede Position Effekte auf der Grundlage aller verwendeten Items und so für jedes Item an jeder Position einen Itempositionseffekt zu berechnen. Dabei muss darauf geachtet werden, dass jedes Item an jeder Position ausreichend häufig beantwortet wurde, um eine angemessene Anzahl an Antworten pro Item zu erreichen (Thompson & Weiss, 2011). Um die Itemparameter möglichst präzise schätzen zu können, sollten mindestens $N = 30$ repräsentative Probanden (besser $N = 100$) auf jedes Item antworten (Johanson & Brooks, 2010). Für eine stabile Schätzung der Itempositionseffekte sollten mindestens $N = 30$ Probanden auf jedes Item an jeder Position antworten. An dieser Stelle wird auf die besondere Bedeutung der Gestaltung des Testheftdesigns und der Prüfung auf Positionseffekte im Zusammenhang mit adaptivem Testen hingewiesen. Beim FIT kann die Gültigkeit der Itemparameter dadurch sichergestellt werden, dass die Items während der Testung an der gleichen Position wie bei der Kalibrierungsstudie vorgegeben werden. Beim adaptiven Testen steht zu Testbeginn jedoch nicht fest, an welcher Stelle welches Item vorgelegt wird. Jedes Item sollte deshalb auf allen möglichen Positionen kalibriert werden.

3.4.2 Itemparameterschätzung, Itemqualität und Modellgültigkeit (inkl. Informationskriterien)

Im Anschluss an die Kalibrierungsstudie erfolgen häufig die Festlegung der Itemparameter, die Prüfung der Itemqualität (Itemselektion) und die Prüfung Modellgültigkeit.

Bei der Prüfung der Itemqualität werden häufig die Itemparameter Schwierigkeit, Trennschärfe (Itemdiskrimination) und Ratewahrscheinlichkeit untersucht. Wie gut die erhobenen Daten das gewählte IRT-Modell abbilden wird über den sogenannten Modellfit untersucht. In diesem Zusammenhang ist es ratsam, die Dimensionalität der Itemantworten zu prüfen. Diese Schritte werden hier unter dem Begriff Pretest eingeordnet. Nachfolgend wird hauptsächlich das Vorgehen für unidimensionale Tests beschrieben. Dennoch ist es ratsam, auch multidimensionale Modelle in Betracht zu ziehen und mit unidimensionalen Modellen zu vergleichen, um die Dimensionalität der Itemantworten zu prüfen. Allgemein ist die Verwendung des sparsamsten Modells, welches dennoch angemessen die Antworten der Probanden abbildet, zu empfehlen. Deshalb sind, insofern theoretische Annahmen nicht dagegen sprechen, bei gleicher Passung unidimensionale Modelle den multidimensionalen Modellen vorzuziehen, da sie weniger Annahmen über das Antwortverhalten machen (Thompson & Weiss, 2011; Wise & Kingsbury, 2000).

Als Methode für die Parameterschätzung im Rahmen der IRT eignet sich u. a. die Maximum-Likelihood-Methode. Die Likelihood-Funktion beschreibt die Wahrscheinlichkeit der beobachteten Testdaten unter der Bedingung des angenommenen Testmodells als eine Funktion von Modellparametern. Es wird zwischen unbedingter (unconditional) maximum likelihood (UML), bedingter (conditional) maximum likelihood (CML) und marginaler maximum likelihood (MML) unterschieden. Die Maxima der drei Funktionen sind Schätzer für die Itemparameter (Rost, 2004). Die MML-Schätzung ist eine häufig verwendete Technik bei der Itemkalibrierung, welche für das 1PL-, 2PL-, 3PL-Modell und auch bei multidimensionalen Modellen funktioniert. Bei Glas (2010) findet sich ein allgemeiner MML-Ansatz für 3PL-Modelle. Generell für das Rasch-Modell ist die Wahrscheinlichkeit einer beobachteten Antwortmatrix $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_U)$ von U unabhängig antwortenden Probanden:

$$P(\boldsymbol{X}|\,\theta, b) = \prod_u \prod_i P(x_{ui}), \tag{4}$$

wobei $\theta_u$und b_i unbekannte fixe Parameter sind. Mit den beobachteten Antworten kann Formel (4) als Likelihood-Funktion für $\theta_u$und b_i gesehen werden und bildet die

Basis für die Itemparameterschätzung (Wainer & Mislevy, 2000). Wenn $p(\theta)$ ein Vorwissen über die Verteilung der Personenfähigkeit (Verteilungsfunktion) ist, dann ist die MML-Schätzung von b:

$$L(b|\,\boldsymbol{X}) = \prod_{u} p(x_i)p(\theta)\,d\theta. \tag{5}$$

Häufig wird für $p(\theta)$ eine Normalverteilungsfunktion angenommen. Aber auch ein simultanes Schätzen der Verteilung ist über das Maximieren der marginalen Wahrscheinlichkeit der beobachteten Antwortmuster (Pattern) möglich (z. B. Eggen & Verhelst, 2011; Wainer & Mislevy, 2000). Numerische Verfahren zur Lösung des Algorithmus können z. B. Quadraturmethoden (z. B. Gauss-Hermite Quadratur) oder Monte Carlo Methoden sein. Als Erweiterung können bei der Schätzung der Itemparameter Posteriori-Verteilungen für die Itemparameter als Information hinzugezogen werden. Dies wird als Bayes Modal Schätzungen (Bayes modal estimates; BME) bezeichnet (Wainer & Mislevy, 2000).

Mit der Schätzung der Itemparameter geht die Wahl des zugrundeliegenden Modells einher. Zur Prüfung der Modellgültigkeit (Modellfit) bietet sich u. a. der Likelihood-Quotienten-Test an. Je höher der Wert L aus der Likelihood-Funktion, desto besser wird das Modell durch die Daten erklärt. Die Likelihood-Funktion im Rasch-Modell ist das Produkt der Patternwahrscheinlichkeiten $P(\boldsymbol{x}_u)$ über alle Personen U:

$$L = \prod_{u} P(\boldsymbol{x}_u). \tag{6}$$

Mit einem Likelihood-Quotienten-Test lassen sich die Ergebnisse aus der Liekelihood-Funktion L von zwei unterschiedlichen Modellen miteinander vergleichen. Hierzu wird die Devianz D, der zweifache negative Logarithmus vom Liekelihoodwert L betrachtet:

$$D = -2\log(L). \tag{7}$$

Mit Hilfe der Anzahl der Parameter der zu vergleichenden Modelle und einem Chi-Quadrat-Differenzentest kann so ebenfalls ein Modellvergleich erfolgen. Des Weiteren können informationstheoretische Maße (Informationskriterien) für die Modellgültigkeitstests genutzt werden. Das *Akaikes information criterion* (AIC) berücksichtigt neben dem Likelihoodwert L zusätzlich die Anzahl an Parametern n_p (Akaike, 1978):

$$AIC = 2(n_p - \log(L)). \tag{8}$$

Das *Bayesian Information Criterion* (BIC) gewichtet die Parameteranzahl stärker mit dem Logarithmus der Stichprobengröße U als das AIC und misst dem Einfachheitskriterium so eine höhere Bedeutung zu (Schwarz, 1978). Dies ist gerade bei großen Datensätzen (also Tests mit vielen Items und Antwortkategorien) sinnvoll:

$$BIC = -2\log(L) + \log(U)\, n_p. \tag{9}$$

Das konsistente (consistent) AIC (CAIC) soll auch bei größerem Stichprobenumfang konsistent bleiben und steht für eine Korrektur des AIC (Bozdogan, 1987):

$$CAIC = -2\log(L) + \log(U)\, n_p + n_p. \tag{10}$$

Rost (2004) gibt den Vorschlag, AIC bei kleinen Itemanzahlen mit großen Patternhäufigkeiten und BIC bei großen Itemanzahlen und kleinen Patternhäufigkeiten als Auswahlkriterium zu nutzen. Mit diesen Informationskriterien können Modelle miteinander verglichen werden, die in keiner hierarchischen Beziehung zueinander stehen. Sie sollten jedoch nicht als alleiniges Auswahlkriterium für ein Testmodell genutzt werden. Nachdem identifiziert wurde, welches Modell am besten zu den vorhandenen Daten passt, kann die Qualität der Items untersucht werden. Häufig wird dabei die Itemtrennschärfe (bzw. Itemdiskrimination) als zentrales Gütekriterium betrachtet (vgl. Kapitel 3.1.1). Die Trennschärfe gibt Auskunft darüber, wie gut ein Item die Personen zwischen z. B. hoher und niedriger Fähigkeit trennt. In der IRT ist die Trennschärfe als Anstieg der Itemfunktion definiert (vgl. 2PL-Modell; Formel (2) auf S. 20). Prinzipiell sollte jedoch die Itemtrennschärfe bei mehrkategoriellen Itemantworten vorsichtig interpretiert

werden. Bei den Indizes zur Prüfung der Itemqualität (Itemfit) kann zwischen residuen-basierten und likelihood-basierten Fit-Maßen unterschieden werden. Residuen-basierte Maße gehen meist von der Differenz der beobachteten Itemantwort x_{ui} und der erwarteten Itemantwort $E(x_{ui})$ aus. Die likelihood-basierten Itemfit-Maße gehen von der Wahrscheinlichkeit des beobachteten Itemvektors $P(\boldsymbol{x_i})$ aus (Knigge, 2011; Orlando & Thissen, 2000; Rost, 2004). Ein Beispiel für ein residuen-basiertes Maß ist die Mean Squared Fit Statistic (MNSQ) bzw. der gewichtete MNSQ (weighted MNSQ; WMNSQ) in der Software ConQuest 3.0.1 (Adams, Wu, Haldane & Sun, 2012). Der MNSQ basiert auf einem standardisierten Vergleich zwischen erwartetem und beobachtetem *Punktwert* (*Score*). Beim Rasch-Modell ist der Erwartungswert für den MNSQ gleich dem Wert 1. Werte nahe 1 stehen somit für eine geringe Abweichung von empirisch beobachteten und erwarteten Itemantworten (Wu, Adams, Wilson & Haldane, 2007). Werte kleiner als 1 können in der Praxis der Testentwicklung meist unproblematisch gesehen werden, da diese Items sinngemäß zu gut zum Modell passen (*Overfit*). Dies ist eine pragmatische Interpretation des Rasch-Modells durch den Autor. Problematisch sind Werte über 1 (*Underfit*), da das Rasch-Modell die Antwortmuster zu schlecht vorhersagt. Für Schulleistungsstudien wie z. B. PISA werden häufig Werte zwischen 0.8 und 1.2 toleriert. Zusätzlich kann der Itemfit inferenzstatistisch überprüft werden, indem der in ConQuest zugehörige t-Wert als Prüfwert genutzt wird. Bei einer Irrtumswahrscheinlichkeit von 5 %, lägen die t-Werte außerhalb des Intervalls $[-1.96, 1.96]$ (Knigge, 2011; Orlando & Thissen, 2000). Eine weitere Möglichkeit, Itemmisfit zu evaluieren, ist es, die empirische Antwortkurve mit der theoretischen Antwortkurve zu vergleichen (Wise & Kingsbury, 2000). Weitere Modell- bzw. Itemannahmen, die geprüft werden können, sind die Ratewahrscheinlichkeit (vgl. 3 PL-Modell; Formel (3) auf S. 20) oder der *Speededness* eines Tests (Hambleton & Swaminathan, 1985). Neben den genannten Ausschlusskriterien (z. B. Trennschärfe) können weitere Kriterien bei der Itemselektion berücksichtigt werden. Solch ein Kriterium ist Differential Item Functioning.

3.4.3 Differential Item Functioning (DIF)

Wenn Probanden unterschiedlicher Gruppen (z. B. Gruppenzugehörigkeit nach Geschlecht) mit derselben latenten Fähigkeit (z. B. Mathematikkompetenz) eine unterschiedliche Wahrscheinlichkeit haben, ein Item korrekt zu beantworten, spricht man von Differential Item Functioning (DIF) im Sinne eines systematischen Effekts (Clauser &

Mazor, 1998). Zumbo (1999) spricht von einem systematischen Fehler, der dazu führt, dass der Test nicht gegenüber allen Personengruppen fair ist. Der systematische Fehler kommt daher, dass die Items Faktoren enthalten, welche für die Messung des eigentlichen Konstrukts irrelevant sind. Innerhalb eines Rasch-Modells weißt ein Item DIF auf, wenn die Lösungswahrscheinlichkeit für ein Item nicht vollständig durch die Fähigkeitsvariable des Probanden und einem fixierten Schwierigkeitsparameter vorhergesagt werden kann (Wu et al., 2007). Dabei ist zu beachten, dass Unterschiede in der Lösungswahrscheinlichkeit auf ein Item nicht immer auf DIF hinweisen müssen, sondern durchaus auf Unterschiede in der latenten Fähigkeit beruhen können. Problematisch ist dies, wenn nach der Kontrolle dieser Gruppenunterschiede die Lösungswahrscheinlichkeit einzelner Items immer noch stark unterschiedlich ist (Holland & Wainer, 1993). Damit diese Unterschiede als Messfehler (Itembias) und somit als DIF interpretiert werden können, dürfen diese Unterschiede nicht auf die unterschiedliche mittlere Testleistung zwischen den Gruppen zurückzuführen sein, sondern müssen auf den Eigenschaften der Items bzw. der Testsituation beruhen. Es gibt unterschiedliche statistische Methoden zur Prüfung nach systematischen Unterschieden in der Lösungswahrscheinlichkeit zwischen zwei oder mehr Gruppen, z. B. Methoden der klassischen Testtheorie wie dem Delta-Plot, Chi-Quadrat-Methoden wie der Mantel-Haenszel Statistik oder IRT-Methoden wie der Multi-Gruppen Modellierung (Clauser & Mazor, 1998; Embretson & Reise, 2000; Wu et al., 2007). Bei der Analyse von DIF im Rahmen der IRT ist es wichtig, dass die Itemparameter in den Gruppen vor einem Vergleich auf dieselbe Metrik gebracht werden. Anschließend können Unterschiede in den Itemparametern z. B. über die ICCs in den Gruppen für ein Item verglichen werden. Als zusätzliche Entscheidungshilfe dienen Schätzungen der Effektgröße und/oder der Signifikanz der unterschiedlich geschätzten Itemparameter in den Gruppen (Clauser & Mazor, 1998; Embretson & Reise, 2000). Die statistische Analyse von DIF im Kontext eines Rasch-Modells im Rahmen der Software ConQuest kann mittels des Multifacetten Rasch-Modells (Linacre, 1994) erfolgen:

$$P(X_{ui} = 1) = \frac{e^{(\theta_u - G_g - b_i + G_g b_i)}}{1 + e^{(\theta_u - G_g - b_i + G_g b_i)}}. \qquad (11)$$

Das Rasch-Modell (vgl. Formel (1) auf S. 18) wird dabei ergänzt durch die mittlere Fähigkeit $G_g$der Gruppe g. Das Produkt $G_g b_i$ spiegelt dabei den Interaktionseffekt zwischen der mittleren Fähigkeit und der Itemschwierigkeit b_i für das Item i wider. In Bezug auf DIF drückt dieser Wert aus, wie unterschiedlich die Wahrscheinlichkeit ausfällt, ein Item korrekt zu beantworten, nachdem die mittleren Kompetenzunterschiede zwischen den Gruppen berücksichtigt bzw. als Haupteffekte herausgerechnet wurden. Weicht $G_g b_i$ für ein Item i signifikant von 0 ab, kann das als ein Hinweis auf DIF für dieses Item gewertet werden (Osterlind & Everson, 2009; Spoden et al., 2015). Solche identifizierten Items sollten anschließend inhaltlich auf DIF geprüft werden. Dazu bieten sich Einschätzungen von Inhaltsexpertinnen und -experten an. Diese können beispielsweise die zuvor statistisch identifizierten Items untersuchen, indem geprüft wird, ob konstruktirrelevante, aber schwierigkeitsbestimmende Itemmerkmale den DIF-Effekt erklären. D. h., es wird gefragt, für welche Gruppe das Item leichter sein kann, wenn der Vorteil, der sich aus der zu messenden Kompetenz ergibt, außer Acht gelassen wird (Holland & Wainer, 1993; Spoden et al., 2015). Ein Beispiel für DIF: In einer Geometrie-Aufgabe soll rechnerisch ermittelt werden, wie weit der Strafstoßpunkt auf einem Fußballfeld von der Torlinie entfernt ist. Die korrekte Antwort ist 11 Meter. Wenn davon auszugehen ist, dass dieses Item SuS der fünften Klasse vorgegeben wird, kann die Annahme getroffen werden, dass Jungen gegenüber Mädchen einen Vorteil haben, da statistisch gesehen mehr Jungen in dem Alter selbst Fußball spielen als Mädchen. Sie könnten damit das korrekte Ergebnis aus Erfahrung wissen. Die eigentlich zu prüfende Geometrie-Kompetenz wäre somit nicht das alleinige Kriterium, was der männlichen Gruppe zum Lösen der Aufgabe hilft. Items, die eindeutig DIF aufweisen, sollten weiter inhaltlich überprüft werden. Grundsätzlich sollte bei der DIF-Analyse zusätzlich zur statistischen Identifikation immer auch eine inhaltliche Analyse, beispielsweise durch Inhaltsexperten, erfolgen (Spoden et al., 2015). Im Zweifelsfall wird empfohlen, auffällige Items aus dem Itempool zu entfernen.

3.4.4 Itempositionseffekte

Nachdem der Itempool bereinigt wurde, können bei Verwendung eines entsprechenden Testheftdesigns die Itempositionseffekte untersucht werden. Bisherige Studien zeigen, dass die Position, an der ein Item vorgelegt wird, Auswirkungen auf die Schwierigkeit des Items bzw. die Leistung der Probanden haben kann (Albano, 2013; Davey &

Lee, 2011; Davis & Ferdous, 2005; Dawis & Whitely, 1976; Eignor & Stocking, 1986; Harris, 1991; Hartig & Buchholz, 2012; Kingston & Dorans, 1984; Kolen & Harris, 1990; Meyers, Miller & Way, 2009; Pommerich & Harris, 2003; Yen, 1980). Bei FIT wird den Problemen, die aufgrund von Itempositionseffekten entstehen können, häufig durch die Verwendung von balancierten Testheftdesigns begegnet. Dabei wird zum Schwierigkeitsparameter jedes Items der Mittelwert aller Positionseffekte addiert (vgl. statistische Kontrolle in Kapitel 3.4.1). Die resultierenden Itemparameter können dann in linearen Testungen mit wenigen Einschränkungen für weitere Analysen genutzt werden. Im Rahmen eines adaptiven Tests werden Itempositionseffekte bisher jedoch selten berücksichtigt. Eine wichtige Annahme beim computerisierten adaptiven Testen ist aber, dass ein Itemparameter über die Positionen, an der das Item vorgelegt wurde, hinweg gleich bleibt. Bei Nichtbetrachtung von Positionseffekten können die Itemparameter verzerrt sein. Verzerrte Itemparameter können zu einer verzerrten Itemauswahl, somit zu einer verringerten Messpräzision und zu einer falschen Schätzung der Personenparameter beim adaptiven Testen führen (Bowles, Wise & Kingsburry, 2008). Es ist somit ratsam, die Itempositionseffekte zu ermitteln.

Dabei gibt es unterschiedliche Möglichkeiten, Positionseffekte zu modellieren. In groß angelegten Vergleichsstudien (z. B. *Programme for International Student Assessment* (PISA) oder *National Educational Panel Study* (NEPS*)*) werden häufig sogenannte Testhefteffekte, die sich auf das gesamte Testheft beziehen, modelliert (OECD, 2009; Pohl & Carstensen, 2012). Diese sind für die Weiterverwendung im adaptiven Test nicht gut geeignet, da dort das verwendete Testheft im Vorhinein nicht feststeht. In verschiedenen Studien wurden Faktorenanalysen (Schweizer, Troche & Rammsayer, 2011; Schweizer, K., Schreiner, M., & Gold, A., 2009) oder Equating-Verfahren eingesetzt (Meyers et al., 2009; Meyers, Murphy, Goodman & Turhan, 2012; Moses, Yang & Wilson, 2007), um den Einfluss der Position auf die Itemparameter zu ermitteln. Diese Methoden werden hier nicht weiter betrachtet. Bedeutsam für die Modellierung von Itempositionseffekten im Rahmen des computerisierten adaptiven Testens erscheinen vor allem Modelle auf Grundlage der IRT. Hier kann zwischen zwei Typen von Modellen unterschieden werden: (a) Modelle mit zufälligen Positionseffekten und (b) Modelle mit fixen Positionseffekten. Modelle mit zufälligen Positionseffekten können u. a. personenspezifische Unterschiede abtragen. Es werden zufällige Positionseffekte auf der Personenseite, als Personeneigen-

schaft, modelliert (Albano, 2013; Debeer & Janssen, 2013; Hartig & Buchholz, 2012). Diese Modelle werden hier ebenfalls nicht weiter berücksichtigt, da die festgesetzten Itemschwierigkeiten innerhalb computerisierter adaptiver Tests personenunspezifisch sein sollten. Es wird die theoretische Annahme getroffen, dass die Itemschwierigkeiten für alle Probanden in der Population gleich sind. Zudem lässt sich ein personenspezifischer Positionseffekt im adaptiven Test z. B. für die Itemauswahl nur schwer nutzen, da die Informationen über die Person erst während der Testung ermittelt werden können (Frey et al., im Druck). Als Modell mit fixen Positionseffekten schlägt Kubinger (2008) das linear-logistische Testmodell (Linear Logistic Test Model, LLTM) vor. Weirich, Hecht und Böhme (2014) zeigen die Nutzbarkeit eines generalisierten LLTM mit einem zusätzlich eingeführten Fehlerterm zur Modellierung fixer Positionseffekte in einem vollständig balanciertem Design. Alexandrowicz und Matschinger (2008) nutzen ein generalisiertes Modell der logistischen Regression mit fixen Itemparametern und vergleichen dieses Modell mit dem LLTM. Am häufigsten werden zur Modellierung fixer Positionseffekte im Rahmen der IRT die vorhandenen logistischen Modelle um eine zusätzliche Facette erweitert, z. B. über das Multifacetten Rasch-Modell (Bowles et al., 2008; Li, Cohen & Shen, 2012). Von den genannten unterschiedlichen Möglichkeiten zur Modellierung eignen sich nicht alle gleichermaßen für die Verwendung in einem adaptiven Test. Empfehlenswert ist es, ein sparsames Modell zu nutzen, welches das Rasch-Modell minimal ergänzt und keine zufälligen Positionseffekte verwendet. Die Nutzung eines Rasch-Modells und die Annahme der Gleichheit der Itemschwierigkeiten für alle Probanden erleichtern (a) die Interpretation der Positionseffekte und (b) die spätere Hinzunahme weiterer Items zum Itempool (inkl. neuer Kalibrierung in einer neuen Population). Zur Modellierung von Positionseffekten unter den genannten Anforderungen und der praktischen Anwendbarkeit innerhalb eines computerisierten adaptiven Tests eignet sich ein Multifacetten Rasch-Modell (Frey et al., im Druck). So kann die zusätzliche Facette *Positionsparameter* später im adaptiven Algorithmus als ein Parameter auf die Itemschwierigkeit addiert werden. Dies wäre bei itemunspezifischen Positionseffekten, also Effekten die für alle Items an einer Position gleich sind, technisch einfach umsetzbar. Anzumerken ist weiterhin, dass sich, exakt betrachtet, im Verlauf eines Tests nicht die Itemschwierigkeit, sondern die Personeneigenschaft ändert. Es wäre damit anzunehmen, dass der Positionseffekt auf der Seite der Personenfähigkeit (Personenparameter) modelliert werden muss. Da CAT jedoch eine gültige Itemschwie-

rigkeit an jeder Position im Test benötigt, wird die Modellierung der Positionseffekte auf Itemseite hier als Hilfsmittel benutzt, um die Itemschwierigkeit und den ungewünschten Positionseffekt zu bereinigen. So kann die Kompetenz einer Person unabhängig von unerwünschten Positionseffekten auf Personenseite ermittelt werden. Konkret wird an dieser Stelle zur Modellierung von Itempositionseffekten im Rahmen des Rasch-Modells ein 3-Facetten-Rasch-Modell, mit den Facetten Fähigkeit der Person, Itemschwierigkeit und Effekt der Position, vorgeschlagen:

$$P(X_{upi} = 1,) = \frac{e^{(\theta_u - P_p - b_i)}}{1 + e^{(\theta_u - P_p - b_i)}}. \quad (12)$$

P ist dabei die Wahrscheinlichkeit, dass eine Person u ein Item i auf Position p korrekt beantwortet, wobei b_i die Itemschwierigkeit für Item i, θ_u die Fähigkeit der Person u und P_p der Effekt der Position p ist. Der Itempositionseffekt wird hier als Variation der Itemschwierigkeit in Abhängigkeit von der Position eines Items innerhalb eines Testhefts definiert (Leary & Dorans, 1985). Soll der Itempositionseffekt itemspezifisch sein, wird Formel (12) um eine weitere Facette erweitert:

$$P(X_{upi} = 1) = \frac{e^{(\theta_u - P_p - b_i - \delta_{ip})}}{1 + e^{(\theta_u - P_p - b_i - \delta_{ip})}}. \quad (13)$$

Die Facette δ_{ip} bildet dabei die Schwierigkeit von Item i auf der Position p ab. Mit diesem im Sinne eines Interaktionsterms zu verstehenden Parameter wird systematische Varianz zur Vorhersage der Lösungswahrscheinlichkeit von Item i modelliert, die über die Schwierigkeit b_i des Items und den Effekt der Position P_p hinausgeht. Zwischen den beiden vorgeschlagenen Modellen sind weitere Modelle mit Abstufung der Komplexität denkbar, bei denen die Positionseffekte nicht für alle Items gleich, aber auch nicht für jedes Item unterschiedlich sind. Frey et al. (im Druck) nutzt beispielsweise ein Modell, welches die Möglichkeit einräumt, Positionsparameter für unterschiedliche Gruppen von Items (z. B. unterschiedliche lange Items; unterschiedlicher Antwortmodus zwischen den Items) zu untersuchen.

3.4.5 Zusammenfassung

In diesem Kapitel wurde über die Notwendigkeit und die Schritte eines Pretests geschrieben. Dabei wurde auf die Kalibrierungsstudie und die Notwendigkeit eines Testheftdesigns hingewiesen. Besonderes Augenmerk wurde darauf gelegt, den praktischen Ablauf der Kalibrierungsstudie zu beleuchten. Da innerhalb der Kalibrierung die Festlegung der Itemparameter erfolgt, wurde auf die Itemparameterschätzung, die Prüfung der Itemqualität sowie die Prüfung des Modellfits eingegangen. Dieser Punkt ist besonders hervorzuheben, da die Qualität des Itempools maßgeblich über die Qualität des späteren Tests bestimmt. Im Zusammenhang mit der Itemselektion ist es empfehlenswert, die Items auf DIF zu untersuchen. Dementsprechend wurde eine Methode vorgestellt. Zudem wurde auf den in der Literatur zur praktischen Entwicklung eines adaptiven Tests bisher eher vernachlässigten Punkt der Itempositionseffekte eingegangen. Die Berücksichtigung von Itempositionseffekten beim computerisierten adaptiven Testen ist noch nicht umfassend untersucht. Hier wurde deshalb ein einfaches Modell zur Schätzung von Itemparametern vorgeschlagen. Die daraus gewonnenen Itemparameter und Itempositionseffekte können anschließend im adaptiven Algorithmus einfach berücksichtigt werden.

3.5 CAT – Algorithmus

Nach der Festlegung des initialen Itempools, dem Pretest einschließlich der Itemselektion und der Kalibrierung der Itemparameter, kann der adaptive Algorithmus festgelegt werden. Dabei gibt es unterschiedliche Möglichkeiten, den Algorithmus anzupassen. In der nachfolgenden Darstellung wird sich auf die Festlegung des Startpunktes (des vorläufigen Personenparameterschätzers und der Itemauswahl zu Beginn der Testung), die Itemauswahl (während der Testung), die Fähigkeitsschätzung und das Testende (Abbruchkriterien) beschränkt. Zudem wird auf zusätzliche Restriktionen bei der Itemauswahl eingegangen, wie das Ausbalancieren der Inhaltsbereiche aus dem inhaltlichen Zielkonstrukt (Constraint-Management/Content-Balancing) oder die Kontrolle der Häufigkeit der Vorgabe von Items (Exposure-Control). Ein einfacher Algorithmus für CAT ist z. B. bei Linacre (2000) zu finden. Die folgende Abbildung 2 zeigt ein mögliches Flussdiagramm für einen computerbasierten maßgeschneiderten

adaptiven Algorithmus. Die einzelnen Schritte des Flussdiagramms werden nachfolgend genauer erläutert.

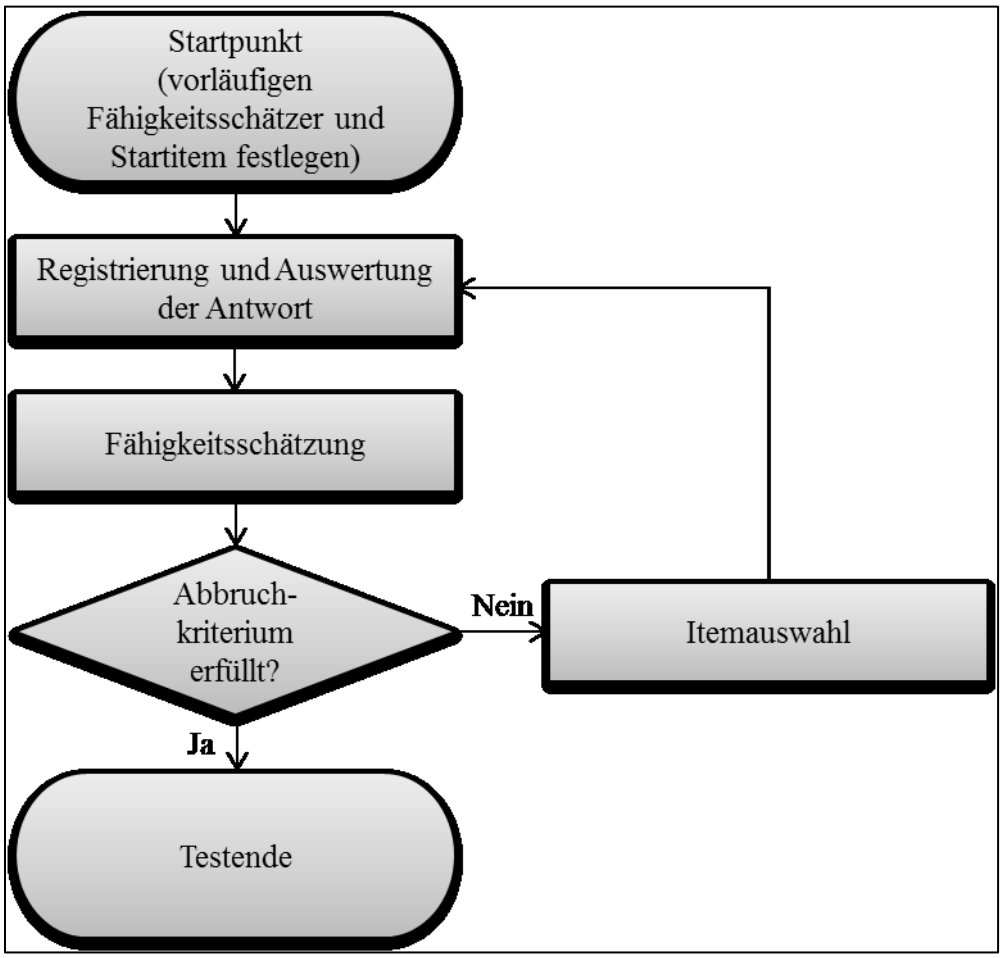

Abbildung 2. Ablauf eines adaptiven Tests.

3.5.1 Startpunkt

Die Itemauswahl bei einem adaptiven Test orientiert sich am Antwortverhalten des Probanden. Zu Beginn des Tests hat der Proband noch keine Items beantwortet. Deshalb ist es wichtig, einen Startpunkt festzulegen. Der Startpunkt kann gerade bei kurzen Testungen einen hohen Einfluss auf die Messpräzision haben (Frey, 2012). Der Startpunkt bezieht sich in der Regel auf die Itemparameter (hier die Itemschwierigkeit) des oder der Startitems. Als Bezugspunkt dient der vorläufig festgelegte Fähigkeitsschätzer (Personenparameterschätzer). Eventuell vorhandene Informationen über die konkrete Testperson können hinzugezogen werden, um den Startwert (a-priori-Schätzung) des Fähigkeitsschätzers möglichst genau zu wählen. Die Informationen über die Fähigkeiten der Testperson bestimmen dann die Auswahl des Startitems, indem ein Item gewählt

wird, dessen Schwierigkeit der Fähigkeit des Probanden entspricht. Solche Vorinformationen können z. B. Testresultate aus vorherigen Testungen des gleichen oder eines ähnlichen Tests oder aus Messungen von Merkmalen, bei denen ein hoher Zusammenhang mit dem zu messenden Merkmal angenommen wird sein. Wenn keine Vorinformationen vorliegen, wird zu Beginn der Testung häufig ein Item mit mittlerer Schwierigkeit, also einer mittleren Lösungswahrscheinlichkeit für einen durchschnittlichen Probanden (z. B. im Rasch-Modell $P(X_{ui} = 1) = .5$), gewählt. Sinngemäß wird beim Festlegen des Startpunktes somit nicht nur die Schwierigkeit des Startitems, sondern auch der Startpunkt des Personenparameterschätzers festgelegt. Bei der Festlegung, wie ein Startitem gewählt wird, sollten die diagnostische Zielsetzung, die zu untersuchende Stichprobe und die bekannten Vorinformationen über ein Individuum berücksichtigt werden (Frey, 2012). Prinzipiell ist davon auszugehen, dass der Einfluss des Startitems auf die Personenparameterschätzung mit zunehmender Testdauer abnimmt. Jedoch sollte berücksichtigt werden, dass ungünstig gewählte Startitems bei den Probanden zu unerwünschten Effekten wie Angst oder Frustration führen können (Hambleton, Zaal & Pieters, 1991). Teilweise werden deshalb sogenannte *Eisbrecher-Items* mit geringerer Schwierigkeit zu Beginn des Tests genutzt, um den Probanden in den Test einzuführen.

3.5.2 Itemauswahl

Auf Grundlage der Registrierung und der Auswertung einer Antwort auf das erste Item wird eine Schätzung der Fähigkeit des Probanden vorgenommen. In der Regel ist das Abbruchkriterium (vgl. Kapitel 3.5.4) nach der Auswertung des ersten Items noch nicht erreicht, so dass eine Itemauswahl während der Testung stattfinden muss. Vereinfacht dargestellt erfolgt die Itemauswahl so, dass ein Proband als nächstes ein leichteres Item vorgelegt bekommt, wenn er das vorhergehende falsch beantwortet hat und ein schwereres Item, wenn er das vorhergehende Item richtig beantwortet hat. Allgemein kann zwischen zwei- und mehrstufigen Strategien bei der Itemauswahl während der Testung unterschieden werden. Bei der zweistufigen Strategie erfolgt eine einmalige Verzweigung, in dem nach einem kurzen Vortest das Leistungsniveau geschätzt wird und darauf abgestimmt ein längerer zweiter Test vorgegeben wird. Dadurch ist es wenig effizient. Jedoch eignet sich das zweistufige Vorgehen auch für papierbasierte Testungen, da es leicht ohne Computer durchgeführt werden kann. Die mehrstufige Strategie kann in eine fest verzweigte (d. h., vor Testbeginn wird festgelegt, welches Item

bei welchem Antwortverhalten vorgelegt wird) und eine maßgeschneiderte Strategie unterschieden werden. Die mehrstufige maßgeschneiderte Strategie, auch variabel verzweigter Test genannt, ist die heute vorherrschende Form. Sie erlaubt eine feine Anpassung der vorzulegenden Items an das Antwortverhalten, da eine Verzweigung erst während des Tests erfolgt. Dies setzt jedoch die Nutzung eines Computers voraus. Bei der Itemauswahl wird dann das Item gewählt, dass unter der Bedingung der aktuell geschätzten Fähigkeit $\hat{\theta}_i$ optimale Eigenschaften aufweist. Zwei Ansätze zur Itemauswahl werden bei mehrstufigen maßgeschneiderten adaptiven Tests verwendet: die Itemauswahl nach Iteminformation und die Itemauswahl nach dem Bayes-Ansatz (Frey, 2012). Dabei berücksichtigt der Bayes-Ansatz die zu Beginn der Testung vorliegenden a-priori-Informationen. Der Ansatz mit Auswahl nach Iteminformation wählt das Item, das bei der momentanen Merkmalsschätzung $\hat{\theta}_u$ den höchsten Wert der Information I, also die maximale Information aufweist (Lord, 1980). Für das Rasch-Modell berechnet sich die Iteminformation I_i für das Item i aus der Multiplikation der Wahrscheinlichkeit das Item i korrekt zu beantworten mit der Wahrscheinlichkeit das Item i nicht korrekt zu beantworten:

$$I_i(\hat{\theta}_u) = P(X_{ui} = 1) * (1 - P(X_{ui} = 1)). \qquad (14)$$

Da der wahre Wert θ üblicherweise nicht bekannt ist, wird zur Berechnung der Iteminformation der vorläufige Fähigkeitsschätzer $\hat{\theta}_u$ zum aktuellen Testzeitpunkt verwendet.

3.5.3 Fähigkeitsschätzung/ Personenparameterschätzung

Es gibt verschiedene Methoden zur Schätzung von Personenparametern: Maximum Likelihood Estimation (MLE), Weighted maximum Likelihood Estimation (WLE), bayesian Expected A Posteriori estimation (EAP), bayesian Maximum A Posteriori estimation (MAP) oder Bayes Modal Estimation (BME). Ausführliche Erläuterungen dazu finden sich u. a. bei Embretson und Reise (2000) oder Hambleton und Swaminathan (1985). Die MLE ist eine asymptotisch erwartungstreue und häufig verwendete Methode, um die Fähigkeit θ einer Person zu schätzen. Wenn $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_i)$ das Antwortmuster nach i Items ist, dann ist die Log-Likelihood-Funktion:

$$\ln(\mathrm{L}(\mathbf{x}|\theta)) = \sum_{i=1} [\, \mathrm{x}_i \ln(\mathrm{P}(\theta)) + (1 - \mathrm{x}_i)\ln(1 - \mathrm{P}(\theta))]. \quad (15)$$

Die beste Schätzung der Fähigkeit kann als Maximum dieser Funktion gefunden werden:

$$\hat{\theta}_{MLE} \equiv \frac{\partial}{\partial \theta} \ln(\mathrm{L}(\mathbf{x}|\theta)) = 0. \quad (16)$$

Die Herleitung des WLE findet sich bei Warm (1989). Alternativ kann anstatt des MLE der BME verwendet werden, welcher beispielsweise in der Software MATE zur Verfügung steht. Dieser kombiniert die Log-Likelihood-Funktion aus Gleichung (15) mit dem Vorwissen in Form einer a-priori-Verteilung $\mathrm{f}(\theta)$. Die a-posteriori Dichtefunktion $\mathrm{f}(\theta|\mathbf{x})$ berechnet sich dann folgendermaßen:

$$\mathrm{f}(\theta|\mathbf{x}) = L(\mathbf{x}|\theta) \frac{\mathrm{f}(\theta)}{\mathrm{f}(\mathbf{x})}. \quad (17)$$

Die a-priori-Verteilung $\mathrm{f}(\theta)$ wird als eine normalverteilte Dichtefunktion angesehen:

$$\mathrm{f}(\theta) = L(\mathbf{x}|\theta) \frac{1}{\sigma\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \quad (18)$$

Um den BME zu berechnen, setzt man, analog zum MLE-Ansatz (vgl. Formel (16) auf S. 62), die erste Ableitung der logarithmierten a-posteriori Dichtefunktion auf 0:

$$\hat{\theta}_{BME} \equiv \frac{\partial}{\partial \theta} \ln \mathrm{f}(\theta|\mathbf{x}) = 0. \quad (19)$$

Es besteht die Möglichkeit, dass auf Basis der Log-Likelihood-Funktion die Lösung mit einfachen Umformungen nicht auffindbar ist. In solchen Fällen bedient man sich z. B. numerischen Lösungsverfahren, um die Nullstellen der Log-Likelihood-Funktion zu finden. In der Software MATE wird beispielsweise das Newton-Raphson-Verfahren als numerisches Lösungsverfahren verwendet (Kröhne & Frey, 2013). Sowohl das Maximum-Likelihood- als auch das Bayes-Verfahren eignen sich zur Fähigkeitsschätzung. Die

Likelihood-Verfahren können bei kurzen Tests teilweise nicht angewandt werden, da z. B. bei einem Antwortmuster mit keinem richtig beantworteten Item (alle Antworten falsch) oder bei einem Antwortmuster mit keinem falsch beantworteten Item (alle Antworten richtig), keine Lösung möglich ist. Dies liegt daran, dass die Ableitung der Likelihood-Funktion für die minimal und maximal möglichen Werte unendlich groß bzw. unendlich klein wird. Das Maximum und somit die Fähigkeitsschätzung einer Person wäre somit $-\infty$ bzw. ∞. Bayes-Verfahren haben dieses Problem nicht, neigen aber im Vergleich zu Likelihood-Verfahren zu größeren Standardfehlern (Hambleton et al., 1991). Eine gewichtete ML-Methode beruhend auf dem Bayes-Ansatz wäre der WLE (Warm, 1989). Die Ergebnisse aus einem Logit-Modell werden nachfolgend mit der Einheit Logits angegeben. Durch die Transformation der Ergebnisse in Logits können Werte zwischen minus und plus unendlich angenommen werden.

3.5.4 Testende

Bei der Festlegung des adaptiven Algorithmus ist die Wahl des Testendes von großer Bedeutung. Wenn ein Test zu kurz ist, kann die Personenparameterschätzung ungenau sein. Ist ein Test zu lang, werden unnötig Ressourcen strapaziert und ggf. Iteminhalte preisgegeben. Zudem können bei langen Tests Müdigkeit oder abnehmende Motivation die Testbearbeitung des Probanden beeinflussen (Linacre, 2000). Ein adaptiver Test wird in der Regel dann beendet, wenn ein oder mehrere vorher definierte Abbruchkriterien erreicht werden. Dabei sind unterschiedliche Kriterien und Kombinationen von Kriterien möglich (Hambleton et al., 1991). Ein Test kann beendet werden, sobald:

- alle Items aus dem Itempool vorgegeben wurden,
- eine maximale Anzahl an Items (zuvor festgelegte Testlänge) vorgelegt wurde,
- der Standardfehler der Fähigkeitsschätzung hinreichend klein ist (z. B. $SE(\theta) < 0.2$ Logits),
- die Fähigkeitsschätzung von einer vorher festgelegten Grenze zum Bestehen bzw. Durchfallen weit genug entfernt ist oder
- eine maximale Testzeit erreicht wurde (Frey, 2012; Linacre, 2000).

Der Standardfehler der Fähigkeitsschätzung $SE(\theta)$ wird in dieser Arbeit aus der Testinformation $TI(\theta)$ wie folgt berechnet:

$$SE(\theta_u) = \frac{1}{\sqrt{TI(\theta_u)}}. \quad (20)$$

Die Testinformation $TI(\theta)$ ist dabei die Summe der Iteminformation $I_i(\theta_u)$ von allen vorgelegten Items i (vgl. Formel (14) auf S. 61):

$$TI(\theta_u) = \sum_{i=1} I_i(\theta_u). \quad (21)$$

Dabei ist zu beachten, dass je nach Itempool die Schätzung der Fähigkeit z. B. in den Randbereichen der Fähigkeitsschätzer mit größerem Standardfehler einhergehen kann als in der Mitte der Fähigkeitsverteilung. Hambleton et al. (1991) schlagen vor, in diesem Fall unterschiedliche Level der Messpräzision für verschiedene Fähigkeitsbereiche als Abbruchkriterium anzulegen. Zudem kann eine Testbeendigung unterdrückt werden, solange bestimmte Kriterien wie z. B. eine Mindestanzahl an beantworteten Items nicht erreicht wurde (Linacre, 2000). Je nach den gewählten Abbruchkriterien kann dies zu Unterschieden in der Testbearbeitung bei unterschiedlichen Probanden, z. B. hinsichtlich der vorgelegten Anzahl an Items oder der benötigten Testzeit, führen. Die Wahl des Abbruchkriteriums sollte sich an der Beschaffenheit des Itempools, dem Anwendungskontext und den Rahmenbedingungen der Testdurchführung orientieren. Bei dem Ziel, individuelle Testwerte zu nutzen, ist es ratsam einen vergleichbaren Standardfehler über die Fähigkeitsschätzer zu erhalten. Bei großangelegten Vergleichsstudien, wo Gruppenergebnisse im Fokus liegen und die Rahmenbedingungen meist sehr restriktiv sind, bietet sich eine Kombination aus maximaler Itemanzahl oder minimalen Standardfehler und Testzeitbeschränkung an (Frey, 2012).

Um einen Anhaltspunkt zu erhalten, welche maximale Itemanzahl vorgelegt werden soll, kann die Reliabilität des Tests durch vorab durchgeführte Simulationsstudien für unterschiedliche Testlängen berechnet werden (vgl. Kapitel 3.2.2). Nachfolgend wird die häufig verwendete Idee der quadrierten Korrelationen von wahren und geschätzten Werten für θ als Reliabilität bei Simulationsstudien dargestellt (Kim, 2012). Das Reliabilitätsmaß $\rho(\hat{\theta}\theta)^2$ berechnet sich aus der quadrierten Korrelation zwischen dem geschätzten Theta $\hat{\theta}$ und dem wahren Theta θ. Es ergibt sich aus dem quadrierten Quotienten der Kovarianz des geschätzten und des wahren Thetas $\sigma(\hat{\theta}\theta)$ und dem

Produkt der Standardabweichungen des geschätzten Thetas $\sigma(\hat{\theta})$ mit der Standardabweichung des wahren Thetas $\sigma(\theta)$:

$$\rho(\hat{\theta}\theta)^2 = r_{\hat{\theta}\theta}{}^2 = \left(\frac{\sigma(\hat{\theta}\theta)}{\sigma(\hat{\theta}) * \sigma(\theta)}\right)^2. \tag{22}$$

Bei der Berechnung der Reliabilität aus empirischen Daten wird eine ähnliche Gleichung verwendet. Da das wahre Theta θ jedoch nicht bekannt ist, wird der Quotient aus der Varianz der Thetaschätzer $\sigma(\hat{\theta})^2$ und der Summe von Varianz der Thetaschätzer mit dem mittleren quadrierten Standardfehler der Thetaschätzer $\frac{1}{N}\sum_{u=1} SE(\hat{\theta})^2$ berechnet. Im Zusammenhang mit adaptivem Testen wird das Reliabilitätsmaß auch als *Fidelity Coefficient* bezeichnet (Kim, 2012).

$$\rho(\hat{\theta}\theta)^2 = \frac{\sigma(\hat{\theta})^2}{\sigma(\hat{\theta})^2 + \varepsilon\,(\sigma(\hat{\theta}|\theta)^2)} = \frac{\sigma(\hat{\theta})^2}{\sigma(\hat{\theta})^2 + \frac{1}{N}\sum_{u=1} SE(\hat{\theta}_u)^2}. \tag{23}$$

Die gleiche Idee liegt auch für die Berechnung der EAP/PV-Reliabilität zugrunde (Adams, 2005). Eine weitere Möglichkeit der Reliabilitätsberechnung ist die Nutzung der sogenannten marginalen Reliabilität (Thissen, 2000). Diese ist in anderer Schreibweise auch unter der Bezeichnung Parallel-Forms Reliability bekannt (Kim, 2012). Sie berechnet sich aus Eins minus des Quotienten des mittleren quadrierten Standardfehlers der Thetaschätzer dividiert durch die Varianz der Thetaschätzer. Bei diesem Reliabilitätsmaß können sich zu Beginn der Testung bei hohen Standardfehlern negative Reliabilitäten ergeben.

3.5.5 Restriktionen

Restriktionen dienen dazu, den Algorithmus an weitere Kriterien anpassen zu können. Die bisher vorgestellten Ansätze zur Itemauswahl während der Testung können beispielsweise dazu führen, dass einige Items sehr vielen Probanden und andere Items nur sehr wenigen bis keinen Probanden vorgelegt werden. Somit steigt die Wahrscheinlichkeit, dass die Inhalte der häufig vorgelegten Items weitergetragen werden. Wenn Iteminhalte Probanden vor der Testung bekannt werden, kann dies die Validität dieser Items in Frage stellen, da z. B. auswendig gelernte Itemantworten nicht mehr zweifelsfrei

auf das zu messende Merkmal zurückgeführt werden können. Unter dem Begriff Exposure-Control werden deshalb Strategien zur Vermeidung unerwünschter Verteilungen der Vorgabehäufigkeiten zusammengefasst. Eine mögliche Strategie ist es, den oben genannten Ansätzen zur Itemauswahl eine stochastische Komponente hinzuzufügen. So kann beispielsweise alternativ zur maximalen Iteminformation eine Bedingung eingeführt werden, dass ein Item per Zufall aus den fünf, acht oder 10 informativsten Items gewählt wird. Die *Sympson-Hetter-Methode*, die *Maximum-Priority-Index-Methode* oder der der sogenannte *Shadow Test* werden häufig zur Exposure-Control genutzt (Frey, 2012).

Inhaltlich ist es häufig auch gewünscht, die theoretische Rahmenkonzeption durch die vorgelegten Items möglichst repräsentativ abzubilden und so zu vermeiden, dass einer Person nur Items bestimmter Teilbereiche der theoretischen Rahmenkonzeption vorgelegt werden. Deshalb ist es hilfreich, den Itemauswahlprozess auch hinsichtlich der inhaltlichen Eigenschaften der Items lenken und optimieren zu können. Hierzu können beim computerisierten adaptiven Testen sogenannte Content-Balancing-Methoden verwendet werden. Um die Anzahl der vorgelegten Items aus jedem Inhaltsbereich (innerhalb einer Kompetenzdimension) kontrollieren zu können, wird im vorliegenden Fall die Methode Maximum-Priority-Index (MPI) beschrieben (Cheng & Chang, 2009). D. h., die Anteile je Inhaltsbereich der betreffenden Domäne werden mit dem MPI angeglichen. Dabei ist $\boldsymbol{C}$ die Constraint-Matrix $I \times K$ mit $c_{ik} = 1$, wenn k ein relevanter Constraint (hier Inhaltsbereich) für Item i ist. Ansonsten ist $c_{ik} = 0$. I ist die absolute Anzahl von Items im Pool und K ist die absolute Anzahl der Inhaltsbereiche. Die Matrix $\boldsymbol{C}$ wird üblicherweise theoretisch durch das inhaltliche Zielkonstrukt bestimmt. Jeder Constraint k hat ein Gewicht w_k. Beispielsweise wird w_k für alle Constraints auf 1 gesetzt, wenn alle Constraints gleich wichtig interpretiert werden sollen. Der Priority Index PI_i für jedes mögliche Item i wird dann berechnet mit:

$$\mathrm{PI}_i = \mathrm{I}_i(\hat{\theta}_u) \prod_{k=1}(w_k * f_k)^{c_{ik}}. \qquad (24)$$

Der erste Teil der Formel gibt die Iteminformation an (vgl. Formel (14) auf S. 61). Diese wird mit dem Produkt $\prod_{k=1}(w_k * f_k)^{c_{jk}}$ gewichtet. Für jeden Constraint repräsen-

tiert f_k dabei die Quote an Items, die für den Constraint k noch nicht vorgegeben wurde und sich folgendermaßen berechnet:

$$f_k = \frac{(X_k - x_k)}{X_k}. \tag{25}$$

Die Variable X_k enthält die Anzahl der insgesamt möglichen Items für einen Constraint k. Die Items, welche bereits aus einem Constraint k vorgelegt wurden, sind in der Variable x_k enthalten. Letztendlich wird immer das Item aus dem Itempool vorgelegt, das den maximalen Priority Index PI_i erzielt.

3.5.6 Zusammenfassung

Es wurde ein Flussdiagramm für einen computerbasierten adaptiven Algorithmus vorgestellt und dessen einzelne Schritte ausführlich erläutert. Bei der Wahl des Startitems gibt es unterschiedliche Möglichkeiten. Falls Vorinformationen über die zu untersuchende Population bzw. des zu untersuchenden Probanden vorliegen, können diese mit genutzt werden. Falls dies nicht der Fall ist, wird häufig ein Item mit mittlerer Lösungswahrscheinlichkeit für einen durchschnittlichen Probanden gewählt. Zudem ist es möglich, leichtere Items als sogenannte Eisbrecher-Items zu nutzen. Neben der Itemauswahl zu Beginn der Testung, ist die Itemauswahl während der Testung ein wichtiger Punkt im adaptiven Algorithmus. Dabei ist das konkrete Vorgehen auch abhängig von der Strategie des Algorithmus (zwei- und mehrstufig sowie fest verzweigt und maßgeschneidert). Die vorherrschende und effizienteste Strategie stellt die mehrstufige maßgeschneiderte Strategie dar. Dazu wurden die Itemauswahl nach Iteminformation und die Itemauswahl nach dem Bayes-Ansatz vorgestellt. Um das zur Personenfähigkeit passende Item wählen zu können, muss eine Methode zur simultanen Schätzung der Personenparameter während der Testung im Algorithmus festgelegt werden. Der MLE und der BME wurden hier genauer betrachtet. Zur Beendigung eines Tests können im Algorithmus mehrere Kriterien (z. B. Anzahl vorgelegter Items, SE, Testzeit) hinterlegt und auch in Kombination verwendet werden. Reliabilitätsanalysen können dazu beitragen, ein angemessenes Testende festzulegen oder die Erfüllung von Restriktionen durch den Algorithmus anzupassen. Dabei wurden zwei Möglichkeiten zur Reliabilitätsmessung (a) für simulierte Daten und (b) für empirische Daten vorgestellt.

Die Wahl des Abbruchkriteriums sollte sich dabei an der Beschaffenheit des Itempools, dem Anwendungskontext und den Rahmenbedingungen der Testdurchführung orientieren. Weiterhin wurde gezeigt, wie über Restriktionen durch Exposure-Control- oder Content-Balancing-Methoden der Algorithmus weiter spezifiziert und angepasst werden kann.

3.6 CAT – Veröffentlichung und Anwendung

Nachdem der Itempool mit festgelegten Itemparametern vorliegt und der Algorithmus festgelegt wurde, können diese Teile in die verwendete Software implementiert werden. Dies geschieht üblicherweise parallel zur Testentwicklung. Erst nach erfolgreicher Implementation der Items in die Software und der Festlegung des Algorithmus kann der Test veröffentlicht und im Feld angewendet werden. Bevor der Test jedoch im endgültigen Anwendungsfeld genutzt wird, sollte eine Pilotierungsstudie durchgeführt werden, in welcher der Algorithmus im Zusammenspiel mit den Items und den Itemparametern empirisch geprüft werden kann. Zudem können die Ergebnisse der Pilotierungsstudie genutzt werden, um die Schätzwerte für die Kompetenzen inhaltlich zu einer aussagekräftigen Skala zusammenzufassen. Um den Test auch nach einer gewissen Zeit noch nutzen zu können, muss der Itempool gepflegt werden. Dies kann u. a. heißen, dass Items abgeändert oder ausgetauscht werden müssen, da sie über die Zeit nicht mehr aktuell sind oder ihren jeweiligen Itemparameter aufgrund unterschiedlichster Einflüsse ändern. In diesem Zusammenhang sollte auch immer wieder die Frage nach der Sicherheit eines Tests gestellt werden. Denn das Bekanntwerden von Iteminhalten über die Zeit verändert häufig die Itemparameter und somit den gesamten Test.

3.6.1 Pilotierungsstudie

Die erste Anwendung des computerisierten adaptiven Tests sollte eine Pilotierungsstudie sein, in der das Zusammenspiel des Itempools, des adaptiven Algorithmus und der Software unter Echtzeitbedingungen getestet wird. Eine wichtige Untersuchung innerhalb der Pilotierungsstudie ist die Prüfung der Simulationsergebnisse aus der Kalibrierungsstudie in Bezug auf den Algorithmus im Zusammenspiel mit den Items (z. B. Reliabilitätsuntersuchungen). Konkret sollte geprüft werden, ob die erwarteten Testlängen, Testzeiten oder Standardfehler aus der Kalibrierungsstudie mit den empiri-

schen Ergebnissen aus der Pilotierungsstudie übereinstimmen. Zudem ist zu prüfen, ob der Algorithmus mit seinen Restriktionen (z. B. Content-Balancing, Exposure-Control) und die Items im Itempool wie gewünscht funktionieren (Thompson & Weiss, 2011). Neben dem Algorithmus wird also auch das Itemmaterial ein weiteres Mal auf seine Qualität geprüft. Die Wahrscheinlichkeit, ein Item korrekt zu beantworten, sollte der festgelegten Lösungswahrscheinlichkeit im Algorithmus entsprechen. Wenn die Wahrscheinlichkeit bei .5 liegt, sollten auch ca. 50 % der Probanden ein Item korrekt beantwortet haben. Dabei sollte die Stichprobengröße, also die Anzahl an Antworten auf ein Item, berücksichtigt werden. Bei Items, die nur sehr selten beantwortet wurden, ist die Varianz meist größer. Somit kann das Ergebnis stark vom Erwartungswert abweichen. Wie groß die Stichprobe insgesamt sein sollte, ist je nach Studie unterschiedlich und kann ebenfalls über eine Simulationsstudie ermittelt werden. Nach Johanson und Brooks (2010) sollte die Anzahl der Probanden der Pilotierungsstichprobe so gewählt werden, dass die Vielzahl der damit verbundenen Aufgabenstellungen berücksichtigt werden. D. h., wenn in der Pilotierungsstudie z. B. gleichzeitig die festgesetzten Itemschwierigkeiten und die Funktionsweise der Itemauswahl geprüft werden sollen, ist die Stichprobengröße anders zu wählen als wenn lediglich eines der beiden Aspekte geprüft wird. In dem hier vorgestellten Vorgehen werden die Itemparameter und die Itemgüte bereits in einer vorherigen Studie, der Kalibrierungsstudie, festgelegt und geprüft (vgl. Kapitel 3.4). Dieser Schritt entfällt somit als Hauptaufgabe der Pilotierungsstudie. In der Pilotierungsstudie, wie sie hier verstanden werden soll, wird deshalb keine Mindestanzahl an Probanden bzw. Antworten auf ein Item erwartet. Insgesamt sollte die Stichprobe der Pilotierungsstudie so gewählt werden, dass sie in relevanten Punkten (z. B. Alter, Geschlecht, Berufsgruppe) gleich zur späteren Zielstichprobe ist. Zudem sollten ausreichend Probanden in den Fähigkeitsbereichen vorhanden sein, die später mit dem Test getestet werden sollen. Bei einer Prüfung von Hochbegabten sollten demnach viele Items im oberen Schwierigkeitsbereich vorhanden sein. Weiter ist es ratsam, die simulierten Reliabilitäten des adaptiven Tests aus der Kalibrierungsstudie an den empirischen Daten der Pilotierungsstudie zu überprüfen. Entsprechend Nutzung der Korrelation als Reliabilitätsmaß bei der Kalibrierung (vgl. Formel (22) auf S. 65) wird hier das als Squared-Correlation Reliability bezeichnete Maß (vgl. Formel (23) auf S. 65) verwendet (Kim, 2012). Die gewonnenen Informationen aus der Pilotierungsstudie dienen im Anschluss dazu, den Algorithmus ggf. anzupassen, weitere Items hinzuzu-

fügen bzw. zu entfernen und die exakten Ergebnisse beispielsweise in einem Manual für die Testanwendung festzuhalten.

3.6.2 Skalenbildung

Die gewonnen Ergebnisse aus einem Test werden zur besseren Interpretation auf ein Maßsystem, der sogenannten Skala abgebildet. Bei der Wahl einer Skala für einen computerisierten adaptiven Test sind die gleichen Faktoren zu berücksichtigen, wie bei der Skalenbildung eines herkömmlichen Tests. Es sollte (a) eine hinreichende Breite der Skala so dass vereinzelte Werte an den Enden der Skala nicht abgeschnitten werden, (b) eine hinreichende Kompaktheit der Skala so dass möglichst wenig Bereiche der Skala ungenutzt bleiben und (c) eine angemessene Zentrierung der Skala so dass der durchschnittliche Punktwert nah bei der Zentrierung der Skala liegt angestrebt werden. Außerdem sollten (d) die Einheiten der Skala der Präzision der Testung angemessen gewählt werden. Bei der Benutzung von Testergebnissen wird häufig angestrebt, die erhobene Leistung der Probanden auf einer inhaltlich gut zu interpretierenden Skala abzubilden. Der Rohsummenwert (Rohscore) eignet sich nicht für CAT. Er ist nicht zwangsläufig über verschiedene Tests mit unterschiedlichen Items hinweg vergleichbar. Er gibt lediglich die Anzahl richtiger Antworten in einem Test zurück und ist somit testspezifisch. Die Schwierigkeit der Items wird dabei nicht berücksichtigt. Dorans (2000) zeigt drei mögliche Punktwerte (Scores), die sich für einen adaptiven Test eignen. Als erstes schlägt er den *Theta-Score* (θ-Score) vor. Die Fähigkeitsskala (Proficiency Scale) auch Theta-Skala (θ-Scale) genannt, lässt sich der IRT zuordnen. Im Rahmen der IRT ist durch die Logitskala eine Möglichkeit geben, die erhobenen Leistungen abzubilden. Tests, deren Items einem IRT-Modell zugeordnet werden, können auf der θ-Skala Punktwerte produzieren. Dies gilt für papierbasierte Testungen ebenso wie für CAT. Allerding benötigt CAT eine Skala, bei welcher der Punktwert der Testung nicht von der eigentlichen Itemauswahl des Tests abhängt. Häufig werden deshalb die θ-Scores bei der Nutzung dieser Skala auf die Standardnormalverteilung transformiert. Als zweiten Score schlägt Dorans (2000) den *Itempoolscore* (IPS) vor. Punktwerte auf der Metrik der θ-Skala können per IRT auf eine andere Metrik gebracht werden. Solch eine Metrik ist z. B. die Itempool-Skala. Der IPS konvertiert über die logistische Funktion für jedes Item den θ-Score in einen Item-True-Score und summiert die Werte dafür über alle Items im Itempool. Der IPS kann als erwartungstreuer Punktwert für einen Probanden

interpretiert werden, wenn der Proband jedes Item im Pool erhalten würde. Als dritten Score schlägt Dorans (2000) den *Item-Subpool-Score* vor. Dieser berechnet für eine Teilmenge der Items im Itempool den IPS. Dieses Vorgehen eignet sich z. B., um Punktwerte des adaptiven Tests mit den Punktwerten eines papierbasierten Tests gleichzusetzen, wenn der papierbasierte Test bereits auf einer gut etablierten Skala verortet ist. Jeder der drei Scores bildet für sich eine eigene Skala. Jedoch können auch eine Vielzahl anderer Skalen aus diesen drei Scores generiert werden. Eine häufig genutzte Skala bei Testungen ist eine Prozentskala von 0 % bis 100 %, die angibt, wie viele Items des gesamten Itempools eine Person korrekt beantwortet hat. Im Idealfall wäre dieser Wert beim adaptiven Test jedoch stets der vorgegebenen Lösungswahrscheinlichkeit. D. h., der Wert läge bei 50 %, wenn die Wahrscheinlichkeit, ein Item korrekt zu beantworten, auf .5 gesetzt wurde. Es ist jedoch möglich, den IPS bzw. den Item-Subpool-Score in die Prozentskala umzuwandeln. Nähere Informationen dazu, zu weiteren Skalen (z. B. der Perzentil-Rang-Skala auf Grundlage des θ-Scores) oder wie man vorhandene Skalen mit einem θ-basierten Punktwert aus einem adaptiven Test ersetzen kann, finden sich bei Dorans (2000).

3.6.3 Erhaltung der Skala

Nachdem die Skala generiert oder sich für die vorhandene θ-Skala entschieden wurde, gilt es, die Skala und somit die Itemparameter aus der Kalibrierungsstudie über die Zeit hinweg zu erhalten. Es gibt unterschiedliche Gründe, warum sich die Itemparameter über die Zeit verändern können. Ein häufiger Grund ist das Bekanntwerden von Iteminhalten. Unterschiede in den Itemparametern zwischen Pretest bzw. Kalibrierungsstudie und der eigentlichen Studie werden oft als Itemparameterdrift bezeichnet. Wenn ein Item von Parameterdrift betroffen ist, bedeutet dies, dass dieses Item später häufiger korrekt bzw. falsch beantwortet wird als noch zu Beginn der Testentwicklung. Die Itemparameter unterscheiden sich dann von den ursprünglichen Parametern. Dies ist ein Grund, Items aus dem Itempool zu eliminieren (Thompson & Weiss, 2011). Ursachen für Itemparameterdrift zwischen Pretest und Haupttestung können darin liegen, dass unterschiedliche Präsentationsmodelle (z. B. computerbasierte Präsentation und papierbasierte Präsentation) verwendet wurden oder aber sich beispielsweise der Lehrplan in der untersuchten Population geändert hat. Aber auch Motivationsunterschiede bei den Probanden über die Zeit sind denkbar. Wenn das Ergebnis eines Tests

die Probanden nicht direkt betrifft oder interessiert (z. B. bei der Kalibrierungsstudie), ist die Motivation vermutlich anders gelagert als wenn der Test beispielsweise über die Abschlussnote entscheidet (Glas, 2010). Das Bekanntwerden von Items über die Zeit kann zudem zu Problemen der Validität des Tests führen. Gerade bei großen Testungen mit vielen Probanden können Iteminhalte schnell öffentlich werden und neuen Probanden bereits vor der Testung bekannt sein. Dies macht es nötig, Items mit bekannt gewordenen Iteminhalten durch neue Items zu ersetzen (Thompson & Weiss, 2011). Deshalb spielt das Thema Testsicherheit eine zentrale Rolle, damit Iteminhalte nicht zu schnell bekannt werden.

Testsicherheit durch Exposure-Control

Testsicherheit hat beim computerisierten adaptiven Testen einen hohen Stellenwert. Denn die Gültigkeit der geschätzten Itemparameter aus der Kalibrierungsstudie hängt mit dem Bekanntwerden der Iteminhalte zusammen. Umso mehr Probanden über den Iteminhalt Bescheid wissen, desto einfacher wird das Item in seiner Itemschwierigkeit. Die Diskrimination des Items geht dann gegen 0 und der Rateparameter wird häufig irrelevant. Für die Itemauswahl und den Scoring-Prozess ist es deshalb wichtig, dass Iteminhalte nicht bekannt werden. Bei häufigen aufeinanderfolgenden Testungen, ist es meist nur eine Zeitfrage, bis die Iteminhalte bekannt werden. Dieses Problem kann sich für computerbasiertes Testen noch verschärfen, wenn in den Testumgebungen (z. B. Schulen) nicht genügend Computer vorhanden sind, um alle Personen parallel zu testen (Wise & Kingsbury, 2000). Um das Bekanntwerden von Items zu reduzieren, gibt es Methoden um die Häufigkeit des Auftauchens von Items zu kontrollieren. Die Kontrolle des Auftauchens von Items wird in der Literatur häufig unter dem Begriff Exposure-Control beschrieben (Glas, 2010). Um Methoden von Exposure-Control anzuwenden, ist es ratsam, einen großen Itempool mit vielen Items in den häufig verwendeten Schwierigkeitsbereichen zu haben. Anderenfalls ist es möglich, dass die Effizienz des adaptiven Algorithmus stark geschwächt wird. Denn wenn Items aus einem Schwierigkeitsbereich häufig genug gezogen wurden, werden diese vorerst gesperrt. Der Algorithmus greift dann auf Items mit der nächsthöheren maximalen Information zurück. Diese Items sind dann häufig weniger informativ als die gesperrten Items. Neben dem Bekanntwerden der Iteminhalte durch Weitersagen ist es im Bildungsbereich oft ein Problem, dass

teilweise auch die Lehrenden bei Bekanntwerden von Testinhalten die SuS auf den Test vorbereiten (*teaching to the test*). Es kann auch vorkommen, dass Testmaterialien, z. B. durch Abfotografieren oder Filmen während der Testung, gestohlen werden. Es gibt zwar verschiedene Möglichkeiten, Itemdiebstahl zu unterbinden (z.B. Colton, 1998), diese sind jedoch häufig sehr kostenintensiv und können ebenfalls keine absolute Testsicherheit gewährleisten. Itemparameterdrift über die Zeit vollständig zu unterbinden, ist deshalb schwer möglich.

Itemparameterdrift

Guo und Wang (2005) zeigen eine Methode, die Skala für CAT stabil zu halten und Itemparameterdrift zu prüfen. Die Größe des Itemparameterdrifts wird evaluiert, indem eine modifizierte quadratische mittlere Abweichung zwischen den Itemparametern verschiedener Studien bzw. Testzeitpunkte ermittelt und diese Differenz anhand von simulierten Werten verglichen wird. Die Daten und Itemparameter aus der ersten Studie bzw. dem ersten Testzeitpunkt (z. B. aus der Kalibrierungsstudie) dienen dabei als Grundlage für die Simulationsstudien. Aus der Stichprobe der ersten Untersuchung werden 10 zufällige Teilstichproben ohne Zurücklegen gezogen. Anschließend werden die Personen- und Itemparameter aus der ursprünglichen ersten Studie genutzt, um für die 10 zufällig gezogenen Teilstichproben neue Antwortvektoren zu erzeugen. Mit Hilfe der neu erzeugten Antwortvektoren werden die Items neu kalibriert und mit der ursprünglichen Kalibrierung verglichen. So kann um die Itemparameter eine Verteilung gelegt werden. Die Itemparameter der zweiten Stichprobe können nun darauf geprüft werden, ob sie innerhalb dieser Verteilung liegen. Glas (2010) zeigt zwei Methoden, um Unterschiede in den Itemparametern, z. B. zwischen Vor- und Haupttest, zu ermitteln. Im Kern wird dort geprüft, ob die Daten beider Testungen dem gleichen IRT-Modell entsprechen. Eine Methode beruht auf einer asymptotischen Testprozedur, die sich auf einen globalen Item-Test stützt, dem *Lagrang-Multiplier-Test*. Die andere Methode zielt auf den Parameterdrift aufgrund des Bekanntwerdens von Items ab. Dabei wird der Annahme gefolgt, dass bekannt gewordene Items über die Zeit leichter werden und nicht mehr so stark diskriminieren. Diese Methode beruht auf einem Instrument aus der statistischen Qualitätskontrolle, der *Cumulative Sum Statistic* und wurde für IRT-Modelle angepasst. Glas (2010) stellt beide praktischen Methoden zum Prüfen von

Itemparameterdrift dar und misst deren Aussagekraft mit Hilfe von Simulationsstudien. Die detaillierten Schritte sind dort nachzulesen.

Um den Itemparameterdrift aufgrund von mangelnder Testsicherheit vorzubeugen, können die Probanden während der Testung genauestens überwacht und Einzeltestungen auf Wunsch vermieden werden (Wise & Kingsbury, 2000). Das Nutzen verschiedener Itempools, die zeitweise rotiert vorgegeben werden, wird in der Literatur ebenfalls empfohlen. Es scheint jedoch sinnvoller, die verschiedenen Itempools zu einem großen Pool zu verknüpfen und damit Exposure-Control-Methoden verwenden zu können. Zusätzlich zum Drift der Itemskala und somit auch zum Drift der Personenskala, kann die Skalenkonsistenz über die Zeit hinweg gefährdet sein. Administrative Änderungen wie z. B. Einführung von Zeitlimits können zu Instabilität der Messskala führen. Diese Instabilität kann durch eine Driftstudie nicht korrekt identifiziert werden. Deshalb müssen relevante Faktoren kontrolliert werden, wenn die Skala über einen langen Zeitraum stabil bleiben soll (Wise & Kingsbury, 2000). Items, die nicht gut zum IRT-Modell passen, neigen zu geringeren Werten für die Itemdiskrimination und erzeugen Fehler bei der Fähigkeitsschätzung. Gerade bei adaptiven Tests, die in der Regel relativ kurz sind, braucht es einen Itempool, der konsistent misst. Möglichkeiten, um Itemmisfit zu prüfen, wurden im Kapitel 3.6.3 vorgestellt. Eine Prozedur zum Identifizieren von schlecht funktionierenden Items sollte deshalb bei der Wartung des Tests und Itempools stets enthalten sein. Identifizierte Items sollten beim Auffinden unverzüglich aus dem Itempool entfernt werden. Es sollten demnach fixe Wartungsintervalle eingeführt werden, in denen die Itemparameter geprüft, Items entfernt und neue Items hinzugefügt werden und der Test auf seine administrativen Aspekte hin überprüft wird.

Entfernen und Hinzufügen von Items

Beim Aufbau und Erhalt eines Itempools sowie der Skala eines adaptiven Tests ist die Identifikation und Elimination schlecht funktionierender Items sehr wichtig. Damit der Itempool nicht immer kleiner wird, müssen neue Items hinzugefügt werden. Üblicherweise werden neue Items vor dem Einpflegen in den Itempool durch einen Pretest geprüft und anschließend oder aber gleichzeitig kalibriert (Thompson & Weiss, 2011). Einem bestehenden Itempool können über Linkingprozeduren Items hinzugefügt werden. Das Linking kann auf viele unterschiedliche Arten erfolgen. Häufig wird in

Vorbereitung auf das Linking einer Gruppe von Probanden der anstehenden Testungen ein Teil der alten bereits kalibrierten Items zusammen mit einem Teil neuer nicht kalibrierter Items vorgegeben. Anschließend wird eine angemessene Linkingprozedur verwendet, um die neuen Items an die bestehende Skala anzubinden. Eine Möglichkeit ist es, zuerst alle (neuen und alten) Items zusammen zu kalibrieren. Anschließend wird die Differenz zwischen den Parametern der alten Items aus der aktuellen Kalibrierung und der ursprünglichen Kalibrierung genutzt, um die neuen Items auf die Originalskala zu transformieren. Dieses Vorgehen ist jedoch suboptimal, da die alten Items bei der aktuellen Kalibrierung im adaptiven Test mitlaufen. Sie werden somit nicht nach einem festen Testheftdesign vorgegeben, wie es bestenfalls in der Kalibrierungsstudie geschehen ist. Die Vergleichbarkeit der aktuellen Skalierungsergebnisse mit der ursprünglichen Skalierung ist somit nicht zwangsläufig gegeben. Für diese Art Linking würde sich eine Kalibrierung mit einem fixierten Testhefts statt eines adaptiven Tests eignen. Das würde wiederum für die Probanden bedeuten, dass sie bei gleicher Messeffizienz längere Tests im Vergleich zu den Probanden mit adaptiven Tests bearbeiten müssen. Eine andere Möglichkeit besteht darin, die neuen Items frei zu schätzen, indem die alten Items auf die Itemparameter fixiert werden, die aus der ursprünglichen Skalierung vorhanden sind. Dabei wird die Fähigkeit einer Person aufgrund der Beantwortung der alten Items bestimmt und aufgrund der Fähigkeit die Schwierigkeit der neuen Items ermittelt (Wise & Kingsbury, 2000). Weitere Methoden für IRT-basiertes Linking finden sich u. a. bei Kolen und Brennan (2014).

3.6.4 Zusammenfassung

Die erste Veröffentlichung und Anwendung computerisierter adaptiver Tests erfolgt häufig in sogenannten Pilotierungsstudien. Unter Echtzeitbedingungen können dort der Algorithmus, der Itempool sowie die simulierten Ergebnisse überprüft werden. Gleichzeitig dient die Pilotierung dazu, das Funktionieren der verwendeten Software im Zusammenspiel mit der verwendeten Hardware im Feld zu erproben. Die Ergebnisse der ersten Studie werden im Rahmen der IRT häufig auf der θ-Skala berichtet. Es ist jedoch ebenso denkbar, eine inhaltlich aussagekräftigere Skala zu nutzen. Solch eine Skala sollte bereits in diesem Schritt geplant werden. Neben der Pilotierungsstudie und der Skalenbildung ist festzuhalten, dass die Anwendung und Veröffentlichung des Tests auch immer beinhaltet, dass die Skala und somit der Itempool gepflegt werden müssen.

Testsicherheit und Itemparameterdrift sind bei der Pflege der Tests wichtige Punkte, die hier beleuchtet wurden. Zur Pflege der Tests gehören ebenfalls die Entfernung von alten und das Hinzufügen von neuen Items durch geeignete Linkingprozeduren. Eine einfache Methode des Linkings wurde beschrieben und auf weitere Möglichkeiten verwiesen. Weitere Hinweise zum Linking finden sich auch im nächsten Kapitel. Zudem wird an dieser Stelle noch einmal darauf verwiesen, dass die Wartung und Pflege des computerisierten adaptiven Tests stets auch die Verwaltung der Testsoftware beinhaltet. Sollen bei einer Wartung Änderungen am Itempool oder am adaptiven Algorithmus erfolgen, bedeutet dies häufig auch Änderungen in und an der verwendeten Software auszuführen. Der Testentwickler sollte deshalb sicherstellen, dass auch nach der Testentwicklung ein Support für die Software besteht oder er selbst die Fertigkeiten und Rechte besitzt, die Änderungen selbstständig vorzunehmen (vgl. Kapitel 3.2.).

3.7 Linking mit papierbasierter Testung

Es kann unterschiedliche Gründe geben, warum unterschiedliche Testarten (z. B. CAT und papierbasiertes FIT) innerhalb einer Erhebung angewendet werden sollen. Ein möglicher Grund ist, dass in manchen Untersuchungsfeldern computerisiertes Testen nicht mit allen Probanden möglich ist. Wenn die Ergebnisse aus den unterschiedlichen Erhebungsinstrumenten anschließend auf derselben Metrik berichtet werden sollen, sind unterschiedliche Aspekte zu berücksichtigen und eine Verbindung zwischen den Punktwerten (Scores) der beiden Testarten vorzunehmen. Eine Verbindung (Linking) zwischen den Punktwerten zweier Tests wird hier definiert als eine Transformation des Punktwertes eines Tests auf den Punktwert des anderen Tests. Dabei gibt es viele unterschiedliche Möglichkeiten des Linkings. Prinzipiell können Linkingprozeduren in drei Bereiche untergliedert werden: *Predicting, Scale Alignment* und *Equating*. Beim Predicting wird aufgrund der Punktwerte von Testart X versucht, die beste Vorhersage für Testart Y, z. B. durch Regressionen, zu treffen. Diese Methode ist im Vergleich zu den anderen beiden Methoden am wenigsten restriktiv und verfolgt das Hauptziel, den Vorhersagefehler möglichst gering zu halten. Predicting ist die älteste Form, um Testwerte verschiedener Tests miteinander zu verbinden. Beim Scale Alignment (kurz Skalierung), besteht das Ziel, den Punktwert von zwei unterschiedlichen Tests auf dieselbe Skala zu transformieren. Die unterschiedlichen Verteilungen der Punktwerte sollen so

zusammengebracht werden. Dabei gibt es unterschiedliche Methoden des Scale Alignment, die je nach Voraussetzung der Situation (z. B. Erhebungsdesign) zu wählen sind. Bei der Wahl der Methode sollte u. a. geprüft werden, ob gleiche oder unterschiedliche Konstrukte miteinander verbunden werden sollen und ob gleiche oder unterschiedliche Populationen zur Skalierung verwendet wurden (Holland, 2007). Die Verwendung unterschiedlicher Konstrukte für unterschiedliche Populationen wird in diesem Zusammenhang auch als *Anchor Scaling* oder Linking mit Ankeritems bezeichnet. Als strengste Form des Linkings kann das Equating gesehen werden. Beim Equating hat ein vorliegender Punktewert dieselbe Bedeutung unabhängig davon, mit welchem Test er ermittelt wurde. Der Zweck des Equating ist es, die Punktwerte zwischen zwei Tests austauschbar zu machen. Das stellt hohe Anforderungen an die beiden Tests und die Equating-Methode. Eine Anforderung ist, dass beide Tests dasselbe Konstrukt auf dem gleichen Schwierigkeitsniveau und mit derselben Reliabilität messen müssen. Es ist nicht immer möglich, alle Voraussetzungen zu erfüllen, um tatsächlich ein Equating vornehmen zu können. Das sogenannte *observed-score test Equating* kann als einfache Adaption des Scale Alignment gesehen werden, um dem Problem des Equating zu begegnen. Detaillierte Informationen zu den unterschiedlichen Möglichkeiten des Linking finden sich u. a. bei Dorans, Pommerich und Holland (2007).

3.7.1 Methoden von Datenerhebungsdesigns

Wie bereits angedeutet, hat das Design, mit dem die Daten erhoben wurden, Einfluss darauf, welche Methoden zum Verbinden von Punktwerten unterschiedlicher Tests angewandt werden können. Die Datenerhebungsmethode ist entscheidend für ein erfolgreiches Linking. Unterschiede in der Verteilung der Antworten über die unterschiedlichen Testformen müssen kontrolliert werden, wenn sie nicht zufällig zustande kommen. Dies wird über sogenannte Datenerhebungsdesigns erreicht. Kolen (2007) nennt drei Faktoren des Datenerhebungsdesigns, die Einfluss auf das Linking haben: Testinhalt (z. B. verwendete Inhaltsbereiche, kognitive Komplexität oder Itemtypen im Test), Messbedingungen (z. B. Testheftdesign, Instruktion, Design der Items, Modus der Testdarbietung) und die untersuchte Population von Probanden (z. B. Geschlecht, Muttersprache, Herkunftsregion, Zeitpunkt zu dem der Test vorgelegt wurde). Wenn diese Faktoren zwischen den zwei zu verbindenden Tests stark abweichen, hat dies natürlich Einfluss auf das Linking. Zwar dürfen die beiden Tests bzw. Testversionen sich

z. B. in ihren Messbedingungen (z. B. computerbasierter und papierbasierter Test) oder in ihrem Inhalt (alte Testversion und neue Testversion) unterscheiden, aber sie müssen stets dasselbe Konstrukt messen, um überhaupt miteinander verbunden werden zu können.

Nachfolgend werden Designs zur Datenerhebung für das Linking besprochen. Häufig genutzt wird das *Zufallsgruppendesign* (Random Groups Design). Hier erhalten z. B. zwei zufällig gewählte Subgruppen Test X oder Test Y. Die Zuweisung erfolgt beispielsweise dadurch zufällig, dass den Schülern im Klassenraum nacheinander abwechselnd Test X und dann Test Y zugewiesen wird. Dieses Vorgehen wird in dieser Arbeit als *spiralisiertes* Vorgeben der Testhefte bezeichnet. Durch diese fortlaufende Zuweisung wird zudem gewährleistet, dass die Testhefte gleich häufig vorgegeben werden. Als weitere Möglichkeit sieht Kolen (2007) das *Einzelgruppendesign* (Single Group Design), in welcher jede Subgruppe beide Testformen aber in unterschiedlicher Reihenfolge erhält. Es liegen somit dieselben Probanden für beide Instrumente vor. Beispielsweise beim Verbinden eines papierbasierten und eines computerbasierten Test, erhält Gruppe A zuerst den papierbasierten und dann den computerbasierten Test und Gruppe B zuerst den computerbasierten und anschließend den papierbasierten Test. Diese Form kann mitunter aber sehr aufwendig sein und für die Teilnehmer teilweise sehr frustrierend, wenn in kurzer Zeit zweimal derselbe Test bearbeitet werden muss. Zudem ist davon auszugehen, dass Reihenfolgeeffekte die Testergebnisse stark beeinflussen. Eine Alternative zum Einzelgruppendesign ist das *Design äquivalenter Gruppen* (*Equivalent Group Design*). Hier bekommen zwei äquivalente Stichproben derselben Population entweder Test X oder Test Y. Die vorgestellten Designs haben strenge Anforderungen an die Daten. Unterschiede in der Verteilung der Probanden stellen die Annahme äquivalenter Gruppen in Frage. Deshalb gibt es Designs mit schwächeren Annahmen. Beispielsweise ein Design mit gleichen Items und nicht äquivalenten Gruppen (Kolen & Brennan, 2014). Dieses Design wird genutzt, wenn lediglich eine Testform pro Testdatum administriert werden kann. In diesem Fall haben beide Testformen ein gemeinsames Set an Items. Die erste Gruppe bekommt Testform X und die zweite Gruppe Testform Y. Hier wird eine systematische Variation zwischen den Testgruppen in Kauf genommen. Die Populationen werden also als nicht äquivalent angesehen. Die auf den gemeinsamen Items geben anschließend eine direkte Information darüber, inwiefern die Leistung der

Probanden zwischen den Gruppen variiert. Dabei muss jedoch sichergestellt werden, dass die gemeinsamen Items in den unterschiedlichen Tests in der gleichen Reihenfolge vorgegeben wurden. Das ist nicht immer möglich. Eine weitere Möglichkeit ist es deshalb, einen zusätzlichen Ankertest bei nicht äquivalenten Gruppendesigns zu nutzen. Hier wird Test X zu der ersten Gruppe und Test Y zu einer zweiten Gruppe zugewiesen. Zusätzlich bekommen alle Gruppen einen identischen Ankertest. Ein Ankerdesign kann jedoch auch folgendermaßen interpretiert werden: Wenn zwei Tests zu zwei Gruppen von Probanden zugewiesen werden, kann ein Anker (a) eine Person sein, die Items von beiden Tests beantwortet hat oder (b) ein Item, welches in beiden Gruppen von Personen vorgelegt wurde (Vale, 1986). Ein Ankeritem wird hier so interpretiert, dass dieses Item in beiden Tests bzw. Testformen vorgelegt wurde. Je nach Design kann somit eine passende Linkingprozedur verwendet werden. Dorans (2000) unterscheidet drei allgemeine Methoden des Equating: equipercentile Methode, lineare Methode und die IRT-basierte Methode. Da die Methoden ein sehr umfangreiches Thema abbilden, wird nachfolgend nur knapp auf die der IRT-basierten Methoden eingegangen. Genaue Angaben zu weiteren Methoden bei entsprechenden Datenerhebungsdesigns finden sich z. B. bei Dorans et al. (2007) oder Kolen und Brennan (2014).

3.7.2 IRT-basierte Methode (Mean/Mean)

IRT-basierte Methoden bieten sich an, da sie die Annahmen der IRT berücksichtigen und die θ-Skala nutzen (Dorans, 2000). Die Lage und Breite einer θ-Skala ist meist unbestimmt. Wenn zwei θ-Skalen zusammengebracht werden sollen, ist es oft notwendig, die eine θ-Skala auf die andere zu transformieren. In manchen Situationen können die beiden Skalen aber auch ohne weitere Transformation zusammengebracht werden (Kolen, 2007). In einem Zufallsgruppendesign können beispielsweise die Itemparameter für einen Test X separat von den Parametern für Test Y geschätzt werden. Denn wenn dieselben Skalierungskonventionen für die Fähigkeiten verwendet werden (z. B. Mittelwert und Standardabweichung einer Standardnormalverteilung für die Verteilung der Personenfähigkeiten), kann angenommen werden, dass die Parameter für die beiden Tests auf derselben Skala liegen. Dann ist keine weitere Transformation notwendig. Bei der Verwendung eines Einzelgruppendesigns können die Parameter für alle Probanden aus beiden Testformen zusammen geschätzt werden, um so die Ergebnisse auf derselben Skala zu berichten (Kolen & Brennan, 2014). Sollten jedoch unterschiedliche Konventio-

nen bei der Skalierung angenommen werden, müssen die Schätzungen der Mittelwerte und Standardabweichungen auf dieselbe Verteilung gebracht werden. Eine typische Situation, in der die Transformation notwendig wird, ist die Verwendung des Designs mit gleichen Items und nicht äquivalenten Gruppen. Hier ist nicht davon auszugehen, dass die Subgruppe von Probanden, welche Test X vorgelegt bekommen hat, in ihrer Parameterschätzung äquivalent zu der Subgruppe mit Test Y ist. Die Parameterschätzer sind somit in der Regel nicht auf derselben Skala. Die Schätzung der Itemparameter für die gemeinsamen Items (Ankeritems) in beiden Tests kann jedoch genutzt werden, um die Transformation der Skala vorzunehmen, z. B. um die gesamte Population zur Schätzung der Items heranzuziehen.

Eine Alternative dazu ist die sogenannte Kalibrierung mit fixierten Parametern (*fixed parameter calibration*), wie sie im empirischen Teil dieser Arbeit angewandt wurde (vgl. Kapitel 4.6). Hier werden die Itemparameter der gemeinsamen Items bei der Skalierung von Testform Y auf die Itemparameter aus der Skalierung von Testform X fixiert. Um die korrekten Werte für die Fixierung zu ermitteln, ist die einfachste Methode bei einem Design mit gleichen Items und nicht äquivalenten Gruppen die Mittelwerte und/oder die Standardabweichung der Itemparameterschätzung der gemeinsamen Items mit einer Mean/Sigma- oder Mean/Mean-Transformation zu ermitteln. Bei der Mean/Sigma-Methode werden die Mittelwerte und Standardabweichungen der b-Parameterschätzer der gemeinsamen Items aus Test X verwendet und für die Skalierung von Test Y darauf fixiert. Bei der Mean/Mean-Methode (Loyd & Hoover, 1980) wird der Mittelwert der a-Parameter der gemeinsamen Items und der Mittelwert des b-Parameters der gemeinsamen Items verwendet. Beim 1PL-Modell, wie es im empirischen Teil verwendet wird, bleibt der Mittelwert des a-Parameters bei eins, wodurch so gesehen nur der Mittelwert des *b*-Parameters bei der Mean/Mean-Methode verwendet wird. Hier wird die Mean/Mean-Methode präferiert. Die nachfolgenden empirischen Analysen werden auf das 1PL-Modell bezogen. Die Schätzung des a-Parameters wird dadurch irrelevant. Bei der Verwendung von 2PL- oder 3PL-Modellen sollten beide Methoden Anwendung finden und verglichen werden (Kolen & Brennan, 2014).

3.7.3 Zusammenfassung

Es gibt unterschiedliche Möglichkeiten, Tests miteinander zu verbinden. Die Methoden des Predicting, Scale Alignment und Equating wurden aufgezeigt. Ausführlicher wurde die IRT-basierte Methode behandelt, bei der die Parameter gemeinsamer Ankeritems der Skalierung der zweiten Testform auf die Mittelwerte der Parameterschätzer aus der ersten Testform fixiert werden (Mean/Mean). Eine direkte Äquivalenz, wie beim Equating angestrebt wird, ist bei der Verwendung zweier unterschiedlicher Testmedien (z. B. computerbasiert und papierbasiert) häufig nur schwer herzustellen. Adaptive Tests sind bei gleicher Reliabilität meist kürzer als papierbasierte Tests. Zudem ist darauf zu achten, dass die Voraussetzungen bei der Testbearbeitung unterschiedlich sein können. Beispielsweise ist das Vor- und Zurückblättern im papierbasierten Test möglich, wohingegen das Weitergehen im Test am PC erst nach der Bearbeitung eines Items möglich gemacht werden kann. Aus diesem Grund wird hier eine separate Kalibrierung des papierbasierten Tests vorgeschlagen, dessen Metrik anschließend mit der Metrik des adaptiven Tests verbunden werden kann. Auf diesem Weg können ungleiche Messbedingungen in die Linkingprozedur einfließen. Wichtig hervorzuheben ist an dieser Stelle noch einmal, dass die zu linkenden Tests bzw. Testformen stets dasselbe Konstrukt messen müssen, um miteinander verbunden werden zu können.

4. Empirische Befunde und praktische Anwendung

In diesem Kapitel werden die vorgeschlagenen theoretischen Schritte zur Testerstellung empirisch am Projekt MaK-adapt angewandt. Die Struktur dieses Kapitels orientiert sich am vorherigen Theorie-Kapitel. Es werden die Schritte zur Testplanung, Entwicklung des initialen Itempools, dem Pretest und der Kalibrierung, dem Algorithmus, der Veröffentlichung und der Anwendung nachvollzogen. Zudem wird eine Linkingprozedur zur Verbindung eines computerisierten adaptiven Tests und eines papierbasierten Tests mit fixer Itemreihenfolge gezeigt. Die Abschnitte sind so gegliedert, dass zuerst Fragestellungen zu den einzelnen Schritten aufgeführt werden, die am Ende des jeweiligen Abschnittes in der Zusammenfassung beantwortet werden. Im Mittelteil der Abschnitte werden die empirischen Ergebnisse aus dem Projekt MaK-adapt zu den einzelnen Schritten dargestellt.

4.1 Testplanung

Im Abschnitt Testplanung wird zuerst das Projekt MaK-adapt vorgestellt. Auf Grundlage des Projektes werden die vorgestellten Schritte zur Erstellung eines computerisierten adaptiven Tests praktisch nachvollzogen und empirisch geprüft. In den Ergebnissen werden die Schritte zur Festlegung des inhaltlichen Zielkonstrukts, die Software und die technische Umsetzung im Projekt MaK-adapt beschrieben.

4.1.1 Fragestellungen

- Was sind die Inhalte und Ziele des Projektes MaK-adapt?
- Welche theoretischen Rahmen wurden gewählt, um die Kompetenzen von SuS in beruflichen Schulen in den drei Domänen Lesen, Mathematik und Naturwissenschaft abzubilden?
- Welche Schritte wurden gewählt, um die inhaltlichen Zielkonstrukte kostengünstig und zeitsparend zu generieren?
- Welche Software wurde für die Entwicklung und Erprobung der Tests verwendet?
- Was sind die Vor- und Nachteile der verwendeten Software z. B. bezüglich Sicherheit und Flexibilität?

- Welche technischen Herausforderungen galt es im Feld der beruflichen Schulen besonders zu berücksichtigen?

4.1.2 Inhalt und Ziele: Projekt MaK-adapt

Ziel des Projektes MaK-adapt war die Entwicklung von drei computerbasierten adaptiven Tests zur Messung der Lesekompetenz, der mathematischen sowie der naturwissenschaftlichen Kompetenz von Berufsschülerinnen und Berufsschülern. Die Messung schulisch erworbener Kompetenzen im berufsbildenden Bereich ist kein Standardvorgehen. Denn Instrumente für eine differenzierte Analyse von Zusammenhängen zwischen allgemeinen Kompetenzen und beruflichen Kompetenzen, welche über das gesamte Leistungsspektrum aller einbezogenen Berufe differenziert, sind dem Autor im deutschsprachigen Raum bisher nicht bekannt. Durch adaptives Testen sollte im Vergleich zum konventionellen, sequentiellen Testen ohne Verlust der Messpräzision sowohl die Testzeit erheblich verringert als auch eine weitgehend konstante Differenzierungsfähigkeit über das gesamte zu erfassende Leistungsspektrum erreicht werden. Das Projekt MaK-adapt hatte somit die Aufgabe, Instrumente zur Erfassung allgemeiner Kompetenzen angemessen auf das heterogene Leistungsspektrum von SuS beruflicher Schulen abzustimmen, so dass die entwickelte Testumgebung allen berufsspezifischen ASCOT-Verbundprojekten für deren Hauptuntersuchung zur Verfügung gestellt werden konnte.

Für das Erreichen des Ziels im neuen Feld war wenig Zeit vorhanden. Laut Projektantrag waren 19 Monate von der Koordination der Projektpartner bis zur Konstruktion der Endversion inklusive dem Verfassen eines Manuals zur Anwendung des Tests geplant. Die Einhaltung der geringen Zeitspanne war notwendig, da die adaptiven Tests durch die anderen Projekte im ASCOT-Verbund bei der Haupterhebung genutzt werden sollten, um schulisch erworbene Kompetenzen als mögliche Determinanten beruflicher Fachkompetenz zu erheben. Um den Zeitplan einhalten zu können, wurde zu Beginn des Projektes MaK-adapt nach bestehenden Kompetenzmodellen und Messinstrumenten von Lesekompetenzen, mathematischen und naturwissenschaftlichen Kompetenzen recherchiert und diese analysiert, um auf vorhandenes Material zurückgreifen zu können. Zudem erfolgten erste Analysen zu den Besonderheiten der Leseanforderungen im beruflichen Kontext. Die Entwicklung der adaptiven Testumgebung in den ersten Monaten des Projektes war nur deshalb möglich, weil eine bereits vorhandene Software zur Erstel-

lung und Administration von adaptiven Tests an die Bedürfnisse von MaK-adapt angepasst wurde. Vor der Durchführung der Kalibrierungsstudie erfolgten die Ausdifferenzierung der Kompetenzmodelle, die Computerisierung der Items, die Rekrutierung der Schulen und die Auslieferung der Tests. Einschließlich der Aufbereitung und der Auswertung der Kalibrierungsdaten waren etwa zehn Monate Zeit veranschlagt. Die restlichen neun Monate wurden zur Erstellung einer vorläufigen computerisierten adaptiven Testform für die drei Domänen, die Pilotierungsstudie, die Aufbereitung und Auswertung der Pilotierungsdaten sowie der Anpassung des adaptiven Algorithmus eingeplant. Anschließend war geplant, für jede Domäne eine Test-Endversion für die ASCOT-Projekte zu erstellen und ein Anwender-Manual zu schreiben. In den weiteren 17 Monaten sollten die Tests in den ASCOT-Projekten angewendet und die Kompetenzniveaus inhaltlich ausdifferenziert werden.

4.1.3 Methode und Ergebnisse: Festlegung inhaltliches Zielkonstrukt

Aufgrund der knappen Zeit und bereits vorhandener theoretischer Zielkonstrukte anderer Studien in den Domänen Lesen, Mathematik und Naturwissenschaft wurde kein gänzlich neuer theoretischer Rahmen konzipiert. Dies war auch nicht notwendig, da es in anderen Studien bereits theoretische Konzepte zur Messung schulisch erworbener Kompetenzen gibt. Deshalb wurden andere Studien wie z. B. *Programme for International Student Assessment* (PISA) oder *Trends in International Mathematics and Science Study* (TIMSS) als Grundlage gesichtet. Die PISA-Studien bieten für alle drei genannten Domänen internationale Testinstrumente an. Zwar ist das Itemmaterial vorhandener Studien nicht passgenau für SuS beruflicher Schulen konzipiert, doch die theoretischen Rahmenkonzepte eignen sich teilweise zur Adaption für die Studien in MaK-adapt. TIMSS untersucht ebenfalls Mathematik- und Naturwissenschaftsleistungen von SuS und hat dementsprechend theoretische Rahmenkonzepte als Grundlage, auf die, im Rahmen des Itemmaterials, zurückgegriffen werden kann.

Für die Domäne Lesen wurde ein theoretisches Zielkonstrukt entworfen, welches die funktionale Lesekompetenz messen möchte (Ziegler, Balkenhol, Keimes & Rexing, 2012). Der funktionale Aspekt ergibt sich aus der Theorie, dass berufliches Lesen zum Großteil *Lesen um zu handeln* ist und dabei andere kognitive Prozesse ablaufen als beim *Lesen um zu lernen*, welches üblicherweise in schulischen Kompetenztests getestet wird. Lesen

wird in dem hier verwendeten theoretischen Konstrukt als Interaktion zwischen dem Leser und dem Text verstanden. Unter dem Begriff Text werden allgemein schriftliche Dokumente, die schriftliche Informationen, Bilder, Diagramme, Tabellen oder andere Arten von Darstellungsformaten enthalten, subsumiert. Die Subdomänen (inhaltliche Dimensionen) werden dabei eingeteilt nach der Format ihrer Repräsentation: (a) deskriptiv (kontinuierliche Texte), (b) hybrid (Mischformen) und (c) depiktional (bildliche Dokumente). Die kognitiven Anforderungen (Leseanforderungen) gliedern sich in Identifizieren, Integrieren und Generieren (Ziegler et al., 2016).

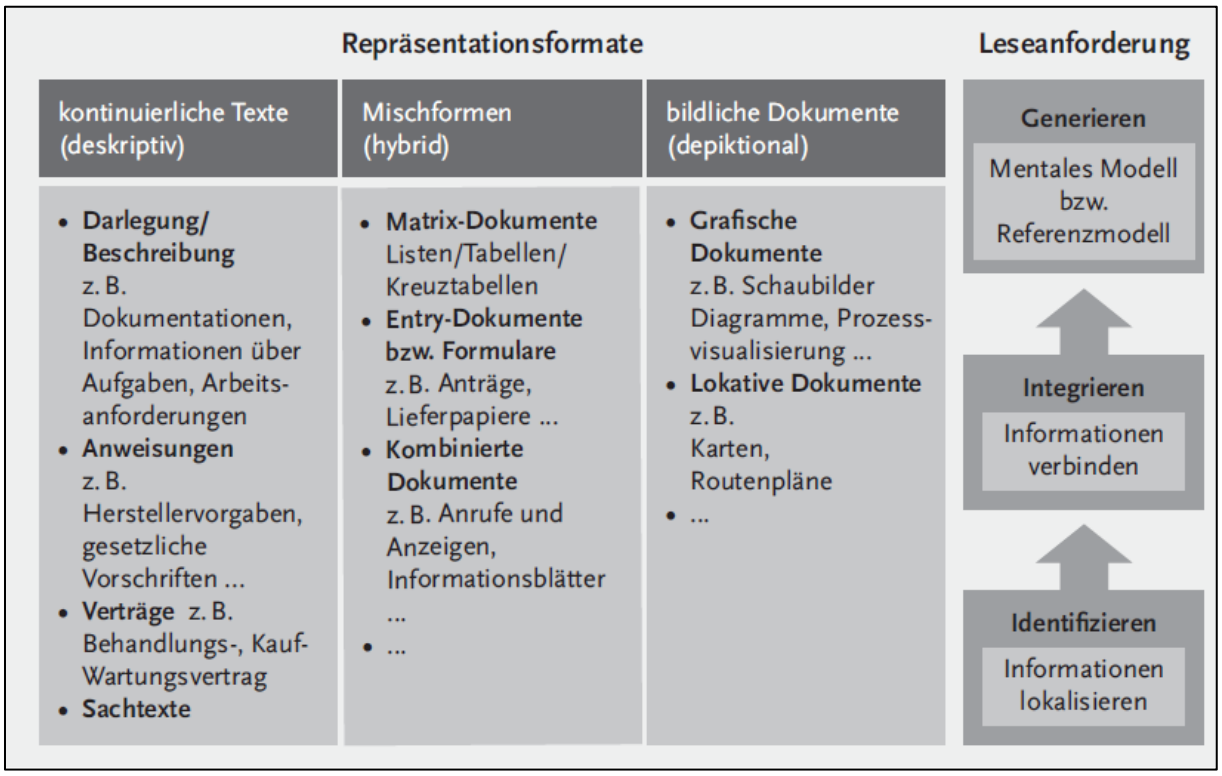

Abbildung 3. Subdomänen und kognitive Anforderungen in der Domäne Lesen (Ziegler et al., 2016).

Zur Auswahl und Klassifikation der Items in der Domäne Mathematik wurde als Grundlage die theoretische Rahmenkonzeption von PISA 2009 (OECD, 2009) genutzt. Die Rahmenkonzeption unterscheidet vier inhaltliche Subdimensionen: (a) Quantität, (b) Veränderung und Beziehung, (c) Raum und Form sowie (d) Unsicherheit und Daten. Zudem wurden bei der mathematischen Kompetenz die drei kognitiven Anforderungen Reproduktion, Verbindung und Reflexion unterschieden. Die im Rahmenkonzept von PISA zusätzlich differenzierten Situationen und Kontexte (z. B. Sport/Gesundheit) wurden bei der Testzusammenstellung im Projekt MaK-adapt nicht berücksichtigt.

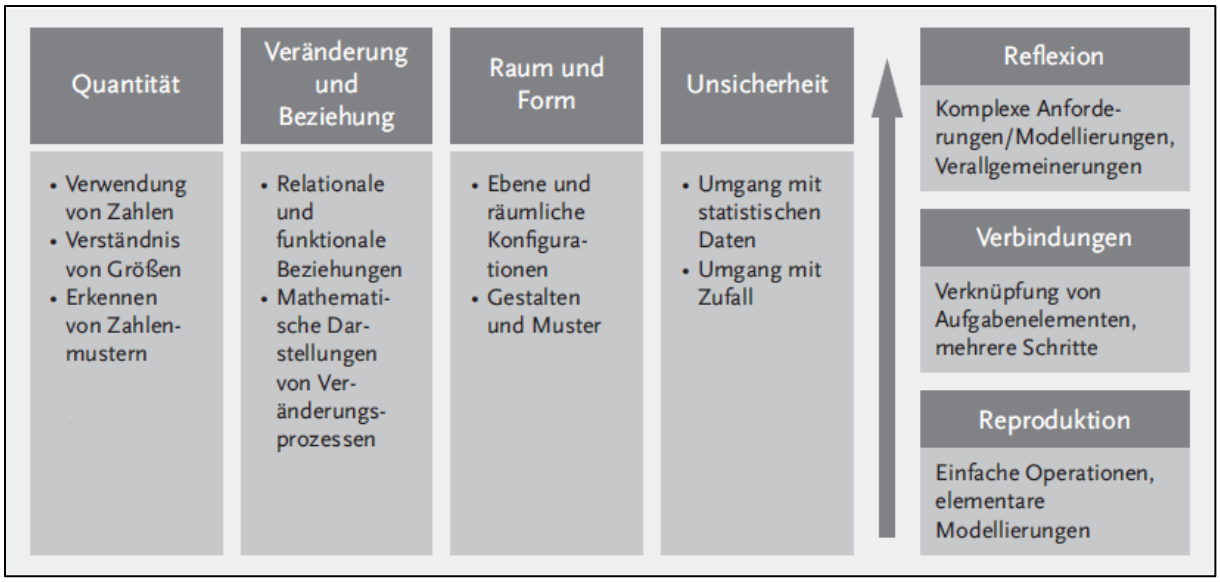

Abbildung 4. Subdomänen und kognitive Anforderungen in der Domäne Mathematik (Ziegler et al., 2016).

Das inhaltliche Zielkonstrukt in der Domäne Naturwissenschaft lehnt sich an den theoretischen Rahmen von TIMSS (Mullis, Martin, Ruddock, O'Sullivan & Preuschoff, 2009) an. Bei der naturwissenschaftlichen Kompetenz werden vier Subdomänen unterschieden: (a) Leben und Gesundheit, (b) Erde, Planeten, Umwelt und natürliche Ressourcen, (c) Stoffe und Stoffveränderungen sowie (d) Bewegung, Kraft und Energie. Diese korrespondieren mit den Inhalten entsprechender Fachgebiete wie Biologie oder Chemie. Die Subdimensionen wurden entsprechend der Fachgebiete strukturiert, um einen besseren Bezug zu beruflichen Anforderungssituationen herzustellen. Berufliche Anforderungssituationen sind häufig fachgebietsübergreifend. Innerhalb der Fachgebiete wird zwischen folgenden kognitiven Anforderungen unterschieden: (a) Verstehen einfacher Informationen, alltagsnahe Schlüsse ziehen, (b) Verknüpfen von Informationen, Bildung einfacher Modelle, (c) Konzeptualisieren, Analysieren und Problemlösen sowie (d) Beherrschen von wissenschaftlichen Verfahren, Umgang mit Theorien.

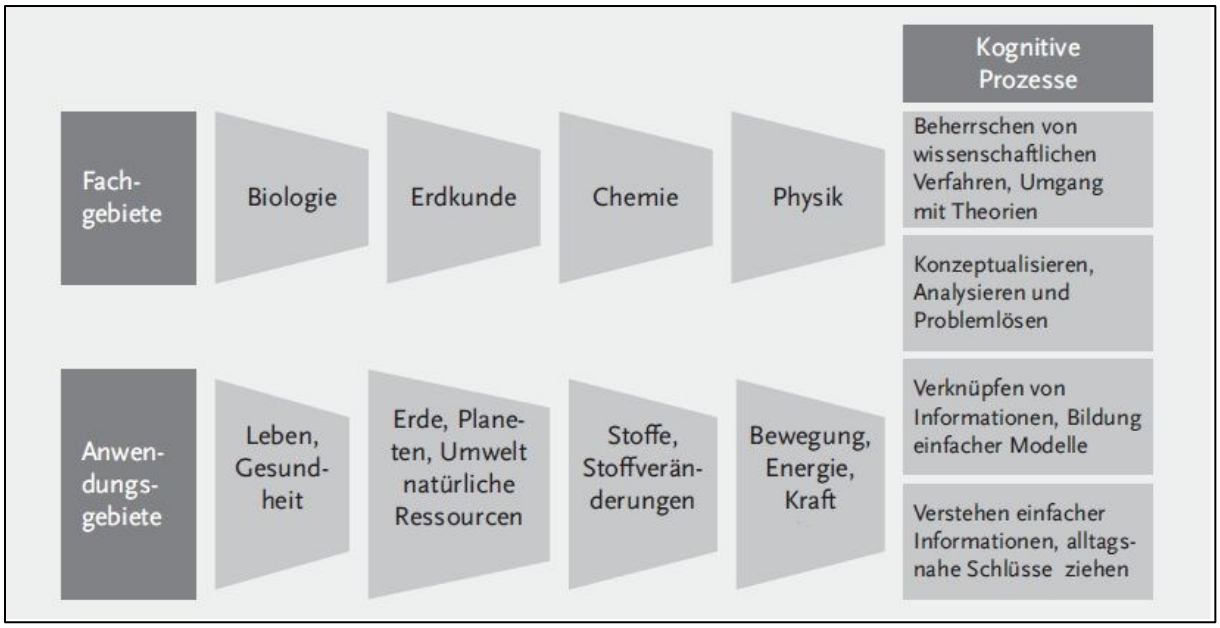

Abbildung 5. Subdomänen und kognitive Anforderungen in der Domäne Naturwissenschaft (Ziegler et al., 2016).

4.1.4 Methode und Ergebnisse: Software und technische Umsetzung

Für die technische Umsetzung der adaptiven Tests wurde im Projekt MaK-adapt die vorhandene Testplattform *Multidimensional Adaptive Testing Environment* (MATE) eingesetzt. MATE wurde im Rahmen des DFG-Projekts *Multidimensionale adaptive Kompetenzdiagnostik* im Schwerpunktprogramm 1293 *Kompetenzmodelle* vom Technology Based Assessment-Cluster am *Deutschen Institut für Internationale Pädagogische Forschung* (DIPF) entwickelt und für das Projekt MaK-adapt angepasst. So konnte nicht nur die lokale MATE-Plattform zur Erstellung, Prüfung und Administration der Items verwendet werden, sondern eine spezielle netzwerkbasierte Lösung Anwendung finden. Ein Vorteil bei der Nutzung von MATE liegt darin, dass in dieser Software direkt Simulationen durchgeführt werden können. Dadurch konnte bei der Erstellung der Tests jederzeit der vorhandene Itempool überprüft werden. Unter anderem wurde geprüft, welche Kombination aus Testlänge und Itemauswahlalgorithmus die höchste Messpräzision für den gegebenen Itempool erwarten lässt. Die Auswertungen der Simulationen können außerdem in der Software MATE direkt graphisch dargestellt werden. Zudem lassen sich die Items direkt in der MATE computerisieren. Neben multidimensionalen und unidimensionalen adaptiven Tests können auch Tests mit fixer Itemreihenfolge erstellt und administriert werden.

Damit der adaptive Algorithmus in der Software MATE entsprechend der Auswahlkriterien auswählen und Bewertungskriterien automatisch bewerten kann, muss ein Itempool mit Antworten hinterlegt sein. Es können unterschiedliche Antwortformate genutzt werden (z. B. Single-Choice bzw. einfache Multiple-Choice, komplexe Multiple-Choice, offene Formate mit eindeutig spezifizierbaren korrekten Antworten). Im Projekt MaK-adapt wurden Single-Choice-Antwortformate (d. h. es gibt vier oder mehr Antwortmöglichkeiten und genau eine davon kann ausgewählt werden und ist korrekt) und offene Textformate mit hinterlegten korrekten Antworten verwendet. Die Items wurden mit Hilfe der Software Microsoft PowerPoint grafisch erstellt (z.B. Festlegung der Anordnung des Itemstamms, der Bilder, der Fragen und der Antworten) und anschließend als Bilddatei über eine sogenannte Schlüsseldatei mit allen weiteren notwendigen Informationen zu den Items (z. B. Itemparameter, Lage der Buttons und Lösungen für offene Items) in die Software MATE eingelesen.

Während der Testplanung wurde sich dafür entschieden, den Itempool und die MATE auf einen lokalen Server an der Friedrich-Schiller-Universität Jena zu hinterlegen. Die erhobenen Daten (Antworten auf die Items, Schätzungen von θ und Log-Daten) wurden direkt nach jedem beantworteten Item auf dem Server gespeichert. Ein Stromausfall bzw. das versehentliche Neustarten eines Computers, an dem getestet wurde, führte somit nicht zum Verlust der Daten. Zudem konnte der Test durch diese Lösung direkt nach dem Neustart an der abgebrochenen Stelle fortgesetzt werden. Die Netzwerklösung wurde gewählt, da die SuS direkt an den beruflichen Schulen getestet werden sollten und vorhandene Rechentechnik (Computerräume) der Schulen genutzt wurde. So wurden keine zusätzlichen Klassensätze von Computern zur Testung in den Schulen benötigt, was mit hohem Transport- und Vorbereitungsaufwand verbunden gewesen wäre. Zur Sicherheit gab es zwei mitgebrachte Laptops, die von den zwei Testleitern vor Ort genutzt werden konnten, falls nicht genügend Rechner vorhanden waren oder eine Schule keinen Computerraum hatte. Der Vorteil der netzwerkbasierten Lösung lag darin, dass theoretisch an jedem Computer zu jeder Zeit eine Testung hätte stattfinden können. Zur Sicherheit wurde der Itempool vor der Testung lokal auf dem Computer abgelegt. Falls es zu einem Netzwerkausfall kommt, kann die Testung mit dem Itempool lokal weiterlaufen und die Daten bleiben so lange lokal gespeichert, bis sie nach einer erfolgreichen Verbindung zum Server automatisch abgerufen wurden. Dieses lokale

Speichern bringt jedoch neue Herausforderungen mit sich. Computer an Schulen sind häufig durch die Sicherheitstechnik stark in Ihrer Benutzung eingeschränkt. Teilweise werden die Administratorrechte ausgelagert, so dass vor Ort niemand einen administrativen Zugang besitzt. Dann ist das lokale Speichern von Daten häufig nicht möglich. Dieser Punkt sollte bei der Testplanung berücksichtigt werden. Deshalb wird eine technische Überprüfung rechtzeitig vor der Testung bei computerisierten Testungen empfohlen. Doch auch ein Installieren des Itempools und das Sicherstellen von Schreibrechten können unzureichend sein, da an öffentlichen Einrichtungen in den Computerräumen häufig am Ende des Tages das System auf ein zuvor gespeichertes Abbild zurückgesetzt wird. Alle vorgenommenen Einstellungen sind danach zurückgesetzt. Zudem hat sich gezeigt, dass auftretende technische Probleme häufig durch konkrete Systemeinstellungen am verwendeten Computer hervorgerufen wurden. Nur selten waren die Systeme aller Computer in einem Computerraum identisch eingestellt. Aufgrund der Nutzung unterschiedlicher Computerräume an unterschiedlichen Schulen ist auch zu erwähnen, dass die Testvoraussetzungen für die SuS sehr heterogen in Bezug auf Mauseinstellung, Bildschirmart, Bildschirmgröße, grafische Darstellung, Lichtverhältnisse usw. waren. Die Software MATE beugt zumindest bei der grafischen Darstellung Problemen vor, da die Iteminhalte automatisch an die Größe des verwendeten Bildschirms angepasst werden, so dass keine Verzerrungen auftreten. Dennoch sollte darauf geachtet werden, dass keine zu kleinen Bildschirme verwendet werden, da sonst einige Inhalte der verwendeten Items nur schwer zu erkennen sind. Die Leistung der verwendeten Computer in dieser Studie ist nachrangig zu betrachten. Es wurden keine aufwendigen Iteminhalte (z. B. Videosequenzen) verwendet und die eigentliche Rechenarbeit der MATE erfolgte auf dem Server. Die Sicherheit der Übertragung wurde dadurch gewährleistet, dass der Itempool als eine passwortgeschützte komprimierte Datei übermittelt wurde. Die Daten der Probanden wurden über ein Hypertext-Transferprotokoll abhörsicher zurück an den Server übertragen. Für die Nutzung eines solchen Protokolls benötigt es ein installiertes Sicherheits-Zertifikat auf dem verwendeten Server. Ist das genutzte Zertifikat in dem verwendeten Browser nicht als vertrauenswürdig eingestuft, kann es zu irritierenden Abfragen kommen und bei fehlenden Administratorenrechten die Testung ggf. daran scheitern. Deshalb sollte zu Beginn geklärt werden, welcher Browser zusammen mit welchem Betriebssystem für die Testung gewählt wird. Möglicherweise ist zusätzlich ein kompatibles Zertifikat zu

erwerben, welches von den Browserherstellern bereits mit der Installation akzeptiert wird. Das erspart unnötige Rückfragen des Browsers über die Vertrauenswürdigkeit der aufgerufenen Seiten. Damit die SuS während der Testungen keine Antworten im Internet des verwendeten Computers suchen konnten, war die Testung so programmiert, dass sich der Browserbildschirm nach Beginn der Testung auf Vollbild stellt und so die Suche über die Suchleiste nicht mehr möglich ist. Mit ein wenig technischem Verständnis oder dem Wissen der Tastenkombination zum Abbrechen ist dieser Schutz durch einen Probanden zwar zu umgehen, stellt aber ein gewisses Hindernis dar. Des Weiteren waren stets mindestens zwei Testleiter vor Ort, um Betrug z. B. durch Abschreiben oder Internetrecherchen vorzubeugen.

Die Navigation zwischen den Items und zwischen den Seiten innerhalb eines Items erfolgt in MATE über Buttons. Ein bereits beantwortetes Item konnte nicht erneut beantwortet werden (Item-Review), da nach der Beantwortung eines Items der Zurück-button ausgeblendet wurde. So konnte der Proband innerhalb eines Items vor und zurück navigieren, nach einem Item jedoch nicht wieder zum Vorherigen zurückgehen. Für das Weiterklicken wurde eine zeitliche Verzögerung programmiert. D. h., der Button, um zum nächsten Item zu gelangen, kann erst nach zwei Sekunden betätigt werden. So wird ein versehentliches Weiterklicken vermieden. Bei der Kalibrierung des Itempools für den adaptiven Test war das Überspringen von Items ohne Beantwortung technisch noch möglich. Dieses Vorgehen produzierte jedoch fehlende Antworten innerhalb eines Antwortvektors eines Probanden und somit zu fehlenden Informationen. Bei der Pilotierung wurde sich deshalb dafür entschlossen, den Weiter-Button zum nächsten Item erst freizugeben, nach dem der Proband eine Antwort gegeben hat. Dies ist zugleich auch eine Interaktion zwischen Proband und Computer. Als fehlende Interaktion in der Software MATE ist anzumerken, dass bei einem offenen Textfeld keine Abfangmethoden bei offensichtlich falschen Eingaben möglich sind (z. B. wenn nach einer Zahl gefragt und die Antwort als Text eingeben wird). Die Interaktion zwischen Computer und Proband erfolgte in den Studien hauptsächlich über eine zu Beginn angezeigte Instruktion. Den Inhalt der Instruktion der Pilotierungsstudie wird nachfolgend wörtlich wiedergegeben.

Liebe Teilnehmerin, lieber Teilnehmer,

vielen Dank für Ihre Bereitschaft an unserer Studie teilzunehmen. Bei dieser werden computerbasierte Testverfahren zur Messung der Kompetenzen von Berufsschülerinnen und Berufsschülern in den Bereichen Mathematik, Lesen und Naturwissenschaften erprobt. Die Tests werden später deutschlandweit an Berufsschulen zur Kompetenzmessung eingesetzt.

Die Teilnahme an der Studie ist freiwillig. Ihre Angaben sind nur Mitarbeiterinnen und Mitarbeitern des Forschungsprojekts „Messung allgemeiner Kompetenzen – adaptiv“ zugänglich, werden ohne Namen gespeichert und nicht an Ihre Schule zurückgemeldet. Die Auswertung der Daten erfolgt anonymisiert. Leistungen einzelner Personen werden nicht ausgewertet.

Die Untersuchung wird insgesamt ca. 90 Minuten dauern. Zu Beginn werden wir Ihnen einige Fragen zu Ihrer Person stellen. Bitte beantworten Sie diese wahrheitsgemäß.

In den darauffolgenden 40 Minuten bekommen Sie Aufgaben aus den Bereichen Mathematik, Lesen oder Naturwissenschaften vorgelegt. Bitte lesen Sie sich die Aufgabenstellung genau durch und klicken Sie danach die Antwort an, die Ihrer Meinung nach richtig ist. Es ist jeweils genau eine Antwort richtig. Bei einigen Aufgaben sind auch Zahlen oder einzelne Wörter einzutragen.

Wichtig zu wissen ist, dass die als nächstes zu bearbeitenden Aufgaben passend zu Ihrer individuellen Leistung im bisherigen Testverlauf ausgewählt werden. Das heißt nach einer von Ihnen getätigten falschen Antwort auf eine Aufgabe bekommen Sie jeweils eine leichtere Aufgabe. Beantworten Sie hingegen eine Aufgabe richtig, bekommen Sie als nächstes eine schwierigere Aufgabe vorgegeben. Dies hat den Vorteil, dass Sie für sich persönlich viel zu einfache oder viel zu schwierige Aufgaben nicht bearbeiten müssen und nur Ihrer Leistung angemessene Aufgaben erhalten. Die Bearbeitungszeit für den Test kann sich deshalb auch stark von der Bearbeitungszeit Ihres Nachbarn unterscheiden.

Infolge des beschriebenen Vorgehens und für den Erfolg der Studie ist es wichtig, dass Sie jede Aufgabe beantworten. Nur so gelangen Sie zur nächsten Aufgabe und können den Test erfolgreich beenden. Außerdem ist zu beachten, dass Sie im Testverlauf nicht zurückgehen können. Sollten Sie eine Aufgabe einmal nicht sicher lösen können, dann geben Sie bitte die Antwort an, die Ihrer Meinung nach am ehesten stimmt.

Einige Aufgaben erstrecken sich über mehrere Bildschirmseiten. Bei solchen Aufgaben können Sie zwischen den einzelnen Seiten mit „Weiter“ und „Zurück“-Buttons (rechts oben) hin und her gehen. Mit einem Klick auf den Button „Nächste Frage“ kommen Sie zur nächsten Testaufgabe. Bitte klicken Sie diesen erst nach der Beantwortung der Frage an, da Sie im Verlauf des Tests nicht mehr zu vorherigen Fragen zurück gehen können.

Die Testleiterin bzw. der Testleiter wird Sie 5 Minuten vor Testende informieren.

Anschließend werden Ihnen weitere Fragen gestellt, die der Beurteilung der Tests und der Testbearbeitung dienen sollen.

Sollten Sie noch Fragen zum Testablauf haben, dann können Sie sich an die Testleiterin bzw. den Testleiter wenden. Dieser wird, sobald alle fertig mit dem Lesen sind, eine entsprechende Frage stellen.

Vielen Dank für Ihre Teilnahme und viel Erfolg!

Anzumerken ist, dass der Proband in der Instruktion auf den Ablauf der adaptiven Testung und das mögliche Gefühlserleben im adaptiven Test hingewiesen wurde. Zudem wurde auf die Tatsache aufmerksam gemacht, dass auf jedes Item eine Antwort gegeben werden muss und dass ein Item-Review nicht möglich ist. Die adaptive Testung wurde zudem auf maximal 40 Minuten beschränkt, was der Tatsache geschuldet ist, dass die ASCOT-Projekte, welche den Test später anwenden sollen, ebenfalls wenig Zeit für die Nutzung dieses Tests zur Verfügung haben. Aus motivationaler Sicht wäre eine Testung ohne Zeitbeschränkung zu bevorzugen. Die restlichen ca. 50 Minuten wurden für die Beantwortung der Fragen zur Person sowie Fragen zur Beurteilung der Tests und der Testbearbeitung verwendet.

4.1.5 Zusammenfassung

In diesem Abschnitt wurde das Projekt MaK-adapt vorgestellt und die Schritte zur (a) Festlegung des inhaltlichen Zielkonstrukts, (b) Wahl der Software und (c) technischen Umsetzung des adaptiven Tests empirisch geprüft sowie am Projekt MaK-adapt beispielhaft nachvollzogen. Ziel des Projektes war es, drei unidimensionale computerisierte adaptive Tests für die Domänen Lesen, Mathematik und Naturwissenschaft zu entwickeln. Diese Tests sollten nach kurzer Zeit in weiteren Projekten im ASCOT-Verbund eingesetzt werden, um effizient schulisch erworbene Kompetenzen mit erheben zu können und so Aufschluss über Zusammenhänge zwischen beruflicher und schulisch erworbener Kompetenz geben zu können. Um computerisierte adaptive Tests in geringer Zeit erstellen zu können, wurde sich bei der Erstellung der inhaltlichen Zielkonstrukte an vorhandene theoretische Rahmen anderer Studien (z. B. PISA, TIMSS) als Grundlage angelehnt. Auf diesem Weg können kostengünstige und in kurzer Zeit erprobte Rahmen genutzt werden. Dies ist möglich, da die verwendeten Studien ebenfalls schulisch erworbene Kompetenzen messen. Die Heterogenität der SuS im beruflichen Kontext wird im darauffolgenden Schritt über die Auswahl der Items berücksichtigt. In der Domäne Lesen stellte sich die Entwicklung etwas aufwendiger dar, da ein Zielkonstrukt zum Messen funktionaler Lesekompetenz (Lesen zum Handeln) entwickelt wurde und die Items bekannter Studien meist das Lesen zum Lernen testen. Dabei spielen häufig andere kognitive Prozesse eine Rolle.

Für die Entwicklung und Erprobung der Tests wurde die vorhandene Software MATE verwendet und angepasst. Die Computerisierung der Items erfolgte über die Software Microsoft PowerPoint. Es wurde kein Item-Review zugelassen und das Weitergehen zum nächsten Item war bei der Pilotierungsstudie erst nach Eingabe einer Antwort möglich. Dieses Vorgehen wurde gewählt, um fehlende Antworten zu vermeiden. Um kostengünstig und zeitsparend testen zu können, wurde sich zur Administration der Tests für eine netzwerkbasierte Lösung entschieden und die Computerräume an den Schulen vor Ort zur Testung genutzt. Die Heterogenität der unterschiedlichen Computer und Netzwerke an den Schulen stellten in einem geringen Teil der Schulen unüberwindbare Hindernisse dar. Fehlende Administratorenrechte, restriktive Firewalleinstellungen oder langsame Internetverbindungen an den Schulen konnten eine Testung teilweise scheitern lassen. Aus diesem Grund wurde sich dazu entschlossen, vor jeder Testung rechtzeitig eine

technische Überprüfung an den Schulen durchzuführen und zu prüfen, ob an jedem Computer der Test durchgeführt werden kann. Das verwendete Testsystem kann als sicher eingestuft werden, da Sicherheitszertifikate und verschlüsselte Ordner für das Verschicken der Daten und Items im Netz verwendet wurden. Die Verwendung einer ausschließlich netzwerkbasierten Lösung und von Computern vor Ort in den Schulen erwies sich als unflexibel. Schulen ohne Computerräume werden so systematisch ausgeschlossen. Aus diesem Grund gab es im Projekt MaK-adapt zusätzlich Laptops, die an die Schulen mitgebracht werden konnten. Für einen standardisierten Einsatz sollten jedoch weitere Auslieferungsmodi ermöglicht werden. Als Alternative wäre z. B. ein Klassensatz Laptops möglich. Die Testungen werden dann lokal auf den Laptops gespeichert. Somit ist die Testung unabhängig von der Stromzufuhr, der Netzwerkarchitektur und der vorhanden Technik an den Schulen. Zudem hätten die SuS stets die gleichen Test-Voraussetzungen, da so sichergestellt werden kann, dass in allen Schulen die gleichen Systeme verwendet werden. Wenn zusätzlich anstatt einer Tastatur am Laptop ein Tablett mit Eingabestift verwendet wird, entspricht das beinahe der ursprünglichen papierbasierten Testung. Nachteilig sind hingegen ein hoher Aufwand für die Verwaltung der Laptops sowie relativ hohe Anschaffungskosten. Zudem muss eine Prozedur implementiert werden, durch welche die Daten nach jeder Testung gesammelt und zusammengefügt werden. Für den Transport eines Klassensatzes Laptops werden voraussichtlich extra Beförderungsmöglichkeiten und mehrere Testleiter benötigt. Zu beachten ist auch, dass die Software auf allen Geräten installiert sein muss. Bei lizensierter Software kann das weitere Kosten verursachen. Insgesamt sind die vorgestellten Schritte zur Testplanung (Entwicklung des theoretischen Zielkonstrukts, Anpassung der Software, technische Umsetzung usw.) in kurzer Zeit und mit wenigen Ressourcen durchführbar. Die Umsetzung der Schritte wurde am Projekt MaK-adapt empirisch geprüft.

4.2 Entwicklung des initialen Itempools

Nachdem der Testentwicklungsprozess zeitlich und finanziell geplant ist, ein inhaltliches Zielkonstrukt definiert wurde und Fragen zu Software und Technik geklärt sind, kann der initiale Itempool entwickelt werden. Die Anforderung, die sich konkret im Projekt MaK-adapt ergibt, ist die Heterogenität der Probanden. Die SuS an den berufli-

chen Schulen können theoretisch das Spektrum von Abgängern ohne Schulabschluss bis hin zu SuS mit Abitur oder sogar Hochschulstudium abbilden. Diese unterschiedlichen Leistungsspektren müssen durch das Itemmaterial abgedeckt sein, um mit wenigen Items möglichst präzise messen zu können. Die Itementwicklung ist im Projekt MaK-adapt aufgrund der geringen Zeit schwierig. Hier wurde sich deshalb auf das Wiederverwenden von bereits bestehenden Items gestützt (Itemrecycling). Nur wenige Items wurden tatsächlich komplett neu entwickelt. Das konkrete Vorgehen wird in dem Ergebnisteil dieses Kapitels beschrieben. Zudem wird auf die Computerisierung der Items im Zusammenhang mit der Software MATE eingegangen. Das Schreiben und Entwickeln von Items sowie das Itempoolmanagement ist in der Software MATE nicht implementiert. Hierzu werden nachfolgend einige praktische Hinweise aufgezeigt.

4.2.1 Fragestellungen

- Wie können in kurzer Zeit mit wenigen Ressourcen gute, für den computerisierten adaptiven Test passende Items generiert werden?
- Wie viel Items müssen im Itempool vorhanden sein, um den adaptiven Algorithmus zu unterstützen?
- Welche Verteilung der Schwierigkeiten der Items wird angestrebt?
- Welches Antwortformat bietet sich an?
- Welche Inhalte (z. B. Bilder, Videos) können in MATE verwendet werden?
- Wie ist eine effektive Computerisierung der Items möglich?
- Wie lässt sich eine Itemdatenbank organisieren?
- Welche Aspekte sind bezüglich Itempoolmanagement und Item-ID zu beachten?

4.2.2 Methode und Ergebnisse: Itemrecycling und Itementwicklung

Aufgrund des Zeitplans wurde sich für das Wiederverwenden von bereits bestehenden Items entschieden. Diese Methode wird hier als Itemrecycling bezeichnet. Dabei wurden vorhandenen Items aus verschiedenen papierbasierten Studien (z. B. PISA; TIMSS; Bildungsstandards; Projekt zur Untersuchung von Leistung, Motivation und Einstellung von SuS berufsbildender Schulen; International Adult Literacy Survey) ausgewählt und die Nutzungsrechte eingeholt. Für die Domäne Lesen wurden die Items

möglichst auf den beruflichen Kontext angepasst. Zur Erstellung des initialen Itempools lagen Items mit den Antwortformaten Single-Choice/einfaches Multiple Choice und komplexes Multiple-Choice sowie offen Items vor. Dabei wurden nur offene Items verwendet, deren Inhalt später automatisch durch den Computer bewertet werden kann (z. B. einzelne Wörter oder Zahlen). In MaK-adapt wurden nur Einzelitems und keine Testlets (vgl. Kapitel 3.3.2) verwendet. Verwertbare Items wurden im ersten Schritt in einer Datenbank gesammelt. Es ist empfehlenswert, zu Beginn der Itemerstellung bzw. -sammlung Überlegungen zu der Itemdatenbank und der Vergabe von Item-IDs anzustellen. Denn bei der Testerstellung eines computerisierten adaptiven Tests werden in der Regel mindestens eine Kalibrierungsstudie und eine Pilotierungsstudie benötigt. Dabei werden häufig über die Studien und die Zeit hinweg Items aus dem Itempool entfernt, geändert oder hinzugefügt. Um dabei die Items stets korrekt zuordnen zu können und einen Überblick zu behalten, sind Itemdatenbanken und Item-IDs unerlässliche Werkzeuge. Im Folgenden wird beispielhaft vorgestellt, auf welchem Weg dies im Projekt MaK-adapt umgesetzt wurde.

Die Itemdatenbank wurde in Microsoft Excel angelegt. Bei größeren Datenbanken empfiehlt es sich ggf. professionelle Produkte zur Erstellung von Datenbanken (z. B. Microsoft Access) zu verwenden. Die Datenbank im Projekt MaK-adapt enthielt für jedes Item Angaben zum originalen Item (Item-ID, Item-Name, Quelle/Studie, Veröffentlichungsdatum, Angaben zum theoretischen Rahmen, Antwortformat, ggf. empirische Schwierigkeiten/Lösungshäufigkeiten in der ursprünglich verwendeten Studie). Die Schwierigkeiten/Lösungshäufigkeiten ermöglichten eine erste theoretische Einordnung der Itemschwierigkeiten. So konnte ansatzweise eingeschätzt werden, ob genügend Items aus verschiedenen Schwierigkeitsbereichen vorliegen. Dies ist möglich, da eine Gleichverteilung der Itemschwierigkeiten angestrebt wurde. Dadurch ist es bei der Itemauswahl möglich, viele Items aus dem entsprechenden Schwierigkeitsbereich passend zur Fähigkeit des Probanden zu ziehen. Dies wiederum erhöht die Messeffizienz. Weiterhin enthielt die Datenbank die projektinterne Item-ID, eine Einschätzung über die Computerisierbarkeit der Items (z. B. Darstellbarkeit im Querformat oder bei offenen Items die Möglichkeit der Bewertung durch den Computer) und bei englischen Items die Übersetzung der Itemnamen ins Deutsche.

Die Item-ID in der Itemdatenbank besteht aus acht Stellen. Die erste Stelle enthält einen Buchstaben, mit dem Hinweis auf die Domäne (M – Mathematik, L – Lesen, N – Naturwissenschaft). Die zweite bis vierte Stelle enthält eine dreistellige Nummer, welche die fortlaufende Itemnummer in der Itemdatenbank darstellt (von 001 bis 999). Die fünfte Stelle gibt einen Hinweis darauf, an welcher Position das verwendete Item im originalen Testlet vorhanden war. Bei der Verwendung von Testlets könnte diese Stelle in der ID auch dazu genutzt werden, die Position des Items im Testlet anzugeben. An der sechsten Stelle ist die Subdomäne des inhaltlichen Zielkonzeptes für die entsprechende Domäne abgebildet. Für die Domäne Mathematik bedeutet 1 – Quantität, 2 – Veränderung und Beziehung, 3 – Raum und Form und 4 – Unsicherheit und Daten. An der siebten Stelle ist die kognitive Anforderung der entsprechenden Domäne laut inhaltlichem Zielkonzept abgebildet. Bei der Domäne Mathematik sind das beispielsweise 1 – Reproduktion, 2 – Verbindungen und 3 – Reflexion. An achter Stelle ist die Versionsnummer des Items bezeichnet. Es ist möglich, dass im Verlauf der Zeit (z. B. nach der Kalibrierungsstudie) Items inhaltlich verändert werden. Um dies in der ID kenntlich zu machen, kann die Item-ID durch die Versionsnummer angepasst werden. Ein Beispiel: Das erste Item in der Datenbank aus der Domäne Mathematik (M001) ist ein Item aus einem PISA-Testlet an erster Position (1), aus der Subdomäne Quantität (1) und mit der kognitiven Anforderung Verbindungen (2). Es ist die erste Version des Items (1) in der Datenbank. Somit erhält das Item die ID M0011121. Die Zuordnung der Items zu dem theoretischen Zielkonstrukt sollte aufgrund der Iteminhalte durch Fachdidaktiker und inhaltliche Experten erfolgen. Das inhaltliche Zielkonstrukt kann nur adäquat durch die Items abgebildet werden, wenn die Zuordnung zweifelsfrei korrekt ist. Für die endgültige Zuordnung der Mathematikitems zum inhaltlichen Zielkonstrukt erfolgte beispielsweise eine Prüfung durch die Abteilung *Didaktik der Mathematik* am *Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik* (IPN) in Kiel.

Bei der Gestaltung der Items wurde darauf geachtet, dass sich an ein zuvor festgelegtes standardisiertes Layout gehalten wurde. Alle Items erhielten einen kurzen Namen und die Item-ID. Anschließend wurde ein ansprechender Stimulus gefolgt von einer Aufgabenstellung dargeboten. Nach der Aufgabenstellung gab es immer eine konkrete Handlungsaufforderung (z. B. *Kreuzen Sie die richtige Lösung an!*). Nach der Handlungsaufforderung folgen bei Single-Choice-Items die Antwortmöglichkeiten und bei offenen

Antworten das Antwortkästchen. Es wurde darauf geachtet, in Sie-Form zu schreiben. Nachdem potentielle Items in der Datenbank gesammelt wurden, kam es zur Endauswahl und Kontrolle der Items. Für die Kontrolle der Items wurden folgende Fragen an das Itemmaterial gestellt:

- Sind die Items inhaltlich korrekt?
- Sind die Iteminhalte für die Zielgruppe geeignet?
- Ist aufgrund des Inhaltes DIF zu erwarten?
- Sind ansprechende und motivierende Stimuli (Bilder, Abhandlungen, usw.) sowie auffordernde Aufgabenformulierungen und sinnvolle Antworten in jedem Item vorhanden?
- Ist die Aufgabe effizient formuliert?
- Sind Rechtschreibung und Grammatik korrekt?
- Wurde sich an ein standardisiertes Layout gehalten (Name, ID, Stimulus, Frage, konkrete Handlungsaufforderung, Antwortmöglichkeit/en, Sie-Form)?
- Sind die als richtig markierten Antworten auch die tatsächlich richtigen Lösungen?
- Sind die Iteminhalte auch auf einem kleinen Bildschirm lesbar?
- Ist die theoretische Zuordnung der Items zum inhaltlichen Zielkonstrukt korrekt?

Die Bearbeitungsdauer der einzelnen Items wurde bereits vor der Kalibrierungsstudie geschätzt. Die Items wurden durch Mitarbeiterinnen und Mitarbeiter an der Friedrich-Schiller-Universität Jena gelöst und die Bearbeitungszeit gestoppt. Die mittlere Bearbeitungszeit diente als erster Anhaltspunkt für die zu erwartende Bearbeitungszeit. Dieses Vorgehen erleichtert das Zusammenstellen Items zu Testheften für die Kalibrierungsstudie (vgl. Kapitel 3.4.1). Wenn bereits ein Itempool vorliegt und Erfahrungen über die Itemzeiten in der zu untersuchenden Population gesammelt wurden, kann ein mathematisches Vorhersagemodell für die Itemzeiten erstellt werden. So können zukünftige Bearbeitungszeiten schneller und genauer geschätzt werden. Eine Möglichkeit wäre es, ein lineares Regressionsmodell aufzustellen, bei dem die abhängige Variable die Bearbeitungszeit ist. Die unabhängigen Variablen könnten z. B. die Anzahl der Wörter im Stimulus, die Länge der Aufgabe, die Anzahl der Antwortmöglichkeiten bei Single-Choice-Items, das Vorhandensein einer offenen Antwortmöglichkeit, die

Anzahl der Grafiken, die Länge eines Medieninhaltes, die Anzahl der Seiten eines Items oder die Anzahl an Tabellen sein. Die Regressionskoeffizienten zeigen dann die zeitliche Abweichung zu einem Item ohne diese Aspekte.

Für einen adaptiven Test ist es sinnvoll, in allen Schwierigkeitsbereichen ausreichend Items im Pool zu haben, um in allen Bereichen der Fähigkeit hinreichend genau differenzieren zu können. Die Kontrolle der Itemschwierigkeiten hilft dabei, (a) nicht zu wenig Items in bestimmten Schwierigkeitsbereichen im Pool zu sammeln und (b) die Testhefte für die Kalibrierung nach Schwierigkeiten etwas auszugleichen. Da im Projekt MaK-adapt gleichverteilte Itempools über die Schwierigkeiten hinweg angestrebt wurden, erfolgte die Zuordnung *theoretischer* Schwierigkeitsparameter zu jedem Item bereits vor der Kalibrierungsstudie. Die Ermittlung der theoretischen Schwierigkeiten erfolgte zum einen durch verschiedene Mitarbeiterinnen und Mitarbeiter an der Friedrich-Schiller-Universität Jena. Zum anderen wurden bei bekannten Lösungshäufigkeiten bzw. Itemschwierigkeiten aus den ursprünglichen Studien, diese Information als Anhaltspunkt für die Schwierigkeit eines Items genutzt. Die empirische Prüfung der theoretisch festgelegten Itemschwierigkeiten durch die Kalibrierungsstudie zeigte jedoch, dass die Abschätzung der Schwierigkeit eines Items, welche für eine andere Population entwickelt wurde, häufig nicht korrekt war. Insgesamt konnten für die Kalibrierungsstudie in der vorhandenen Zeit in der Itemdatenbank für die Domäne Lesen 73 Items sowie die Domänen Mathematik und Naturwissenschaft jeweils 133 Items im initialen Itempool gesammelt werden. In der Domäne Lesen waren nicht so viele Items vorhanden, da einerseits viele Items angepasst oder neu entwickelt werden mussten, so dass sie sich auf die beruflichen Aspekte des Lesens beziehen, und andererseits die Leseitems im Schnitt einen wesentlich längeren Stimulus haben und somit mehr Bearbeitungsdauer für ein Item benötigt wird.

Bei der Zuordnung der Items zum inhaltlichen Zielkonstrukt ging es unter anderem darum, die Subdomänen gleichverteilt abzubilden, da dies später im adaptiven Algorithmus auch bei der Itemauswahl (Content-Balancing) berücksichtigt werden sollte. Die gleichmäßige Verteilung der Items auf die kognitiven Anforderungen wurde vernachlässigt. Grund dafür ist, dass für eine Gleichverteilung der Items nach beiden Dimensionen (inhaltliche Subdomänen und kognitive Anforderungen) wesentlich mehr Items vorhanden sein müssen, um dies über Content-Balancing angemessen abzubilden und

genügend Items mit den entsprechenden Schwierigkeitsparametern im Itempool vorliegen zu haben. Es wurde jedoch darauf geachtet, dass in allen Bereichen kognitiver Anforderungen Items vorhanden sind. Auf diesem Weg, können über die Item-ID im Nachgang beispielsweise Analysen auf Populationsebene zu den kognitiven Anforderungen durchgeführt werden. Es stellte eine Herausforderung dar, in allen Schwierigkeitsbereichen genügend Items zu finden. In der Domäne Mathematik gab es z. B. kaum Single-Choice- oder automatisch zu bewertende offene Items, welche die höchste kognitive Anforderung in der Subdomäne Reflexion messen. In papierbasierten Testungen, aus denen die Items häufig stammen, wurden die Aufgaben meist so formuliert, dass Zeichnungen oder längere Interpretationen als Antwort erwartet wurden. Dies war für die Nutzung der computerisierten adaptiven Tests im Projekt MaK-adapt jedoch nicht möglich, da solch komplexe Antworten nicht automatisch ausgewertet werden konnten. Für Testungen, in denen solche Items dennoch genutzt werden sollen, wird eine Zwischenlösung empfohlen. Es ist möglich, solche Aufgaben im Test mit vorzugeben und die Antworten erst im Nachhinein zu bewerten. Diese Items können dann während des Tests nicht als Information für die Itemauswahl genutzt werden. Die Testzeit wird sich dadurch bei gleicher Messpräzision voraussichtlich verlängern. Im Nachgang können diese Items jedoch wie bei anderen Testungen auch bewertet werden und in die Schätzung der Fähigkeit mit einfließen. Ein Beispiel für die Verteilung der Items in der Domäne Mathematik finden Sie nachfolgend:

Tabelle 2

Verteilung der Items in der Domäne Mathematik (MATH) über die Subdomänen hinweg

	Anzahl Items über Inhaltsbereiche			
Kognitive Anforderungen	Quantität	Veränderung und Beziehung	Raum und Form	Unsicherheit und Daten
Reproduktion	24	11	6	9
Verbindungen	9	18	27	14
Reflexion	3	2	2	8

Ziel bei der Entwicklung des Itempools war es, neben qualitativ hochwertigen Items, welche das inhaltliche Zielkonstrukt adäquat messen, auch genügend Items in den entsprechenden Schwierigkeitskategorien zu haben. Deshalb wurde hier auf die Frage eingegangen: *Wie viele Items sollten in einem Itempool für CAT mindestens vorhanden sein?* Diese Frage kann nicht allgemeingültig beantwortet werden. Der adaptive Algorithmus wird bei der Itemauswahl vor allem dann gut unterstützt, wenn für die entsprechende Fähigkeit eines Probanden genügend Items mit der entsprechenden Schwierigkeit vorliegen. Werden weitere Restriktionen an die Itemauswahlprozedur gestellt (z. B. Content-Balancing oder Exposure-Control; vgl. Kapitel 3.5.5), wird die Anzahl nötiger Items entsprechend höher. Ein Beispiel: Die Verteilung der Itemschwierigkeit und der Personenfähigkeit ist diskret über die fünf Ausprägungen -2, -1, 0, 1 und 2 verteilt. Der Test hat eine Testlänge von 20 Items. Dann enthält der Itempool bestenfalls mindestens 20 Items in jeder der fünf Schwierigkeitsbereiche. Dies entspricht insgesamt 100 Items. Wenn zudem auf Subdomänen Rückmeldung gegeben werden soll, wäre es wünschenswert, in jeder der Subdomänen 100 Items mit der genannten Verteilung zu haben. Dies macht bei vier Subdomänen bereits 400 Items. Da die Skala der Fähigkeiten bzw. Itemschwierigkeiten jedoch stetig und nicht diskret ist, wäre auch ein Vielfaches der genutzten Itemanzahl denkbar. In MaK-adapt wird nicht angestrebt, Rückmeldung auf Subdomänen zu geben. Content-Balancing-Methoden werden lediglich genutzt, um das inhaltliche Zielkonstrukt angemessen abzubilden. Als Größe für den Itempool wurden deshalb 100 Items pro Domäne angestrebt. Beispielsweise wurde in der Domäne Mathematik angestrebt, ca. 20 Items in jedem theoretisch festgelegten Schwierigkeitsbereich (sehr leicht, leicht, durchschnittlich, schwer und sehr schwer) und ca. 25 Items in jeder Subdomäne zu haben. Bei der Planung des Itempools sollte berücksichtigt werden, dass Items nach einer Kalibrierungsstudie z. B. durch Fehler im Item oder aufgrund von Differential-Item-Functioning-Analysen aus dem Itempool ausscheiden können. Es sollten deshalb bis zu 30 % mehr Items in die Kalibrierung genommen werden, als der angestrebte initiale Itempool groß sein soll. Im Projekt MaK-adapt wurden in der Domäne Mathematik für einen angestrebten Itempool von 100 Items deshalb mehr als 130 Items in der Kalibrierungsstudie verwendet. Dabei wurde ebenfalls versucht, diese Items gleichmäßig zu verteilen (ca. 26 Items pro theoretisch festgelegten Schwierigkeitsbereich und ca. 33 Items pro Subdomäne).

Diese Aufgabe stellte sich als schwierig heraus. Häufig gab es gerade in den Randbereichen der Schwierigkeiten (sehr leicht und sehr schwer) nur wenige und in dem Bereich durchschnittlicher Schwierigkeit viele Items. Nachfolgend ist die Tabelle für die Domäne Mathematik beispielhaft abgebildet.

Tabelle 3

Verteilung der Items in der Domäne Mathematik (MATH) über die Inhaltsbereiche und den theoretisch festgelegten Schwierigkeitsbereich

Schwierigkeit	Quantität	Veränderung und Beziehung	Raum und Form	Unsicherheit
Sehr leicht	8	4	5	4
Leicht	10	5	10	6
Durchschnittlich	8	10	9	13
Schwer	6	10	7	6
Sehr schwer	4	2	4	2

4.2.3 Methode und Ergebnisse: Computerisierung der Items

Nach der Erstellung der Itemdatenbank erfolgten die Computerisierung der Items und das Einlesen der computerisierten Items in MATE in mehreren Schritten. Dabei wurde sich an das im Kapitel 4.2.2 beschriebene Layout gehalten. Konkret wurden der Itemname, die Item-ID, der Stimulus, die Aufgabenstellung, die Handlungsaufforderung und die Antwortalternativen immer an derselben Stelle (soweit möglich) angeordnet. So sollte ein unterschiedliches Funktionieren von Items aufgrund ihres Designs vorgebeugt werden. Das Layout und die Einbindung der Items erfolgten über die Software Microsoft PowerPoint. So konnten die verwendeten Stilelemente problemlos grafisch formatiert und auf dem Bildschirm angeordnet werden. Beispielhaft ist in der nachfolgenden Abbildung ein Item abgebildet.

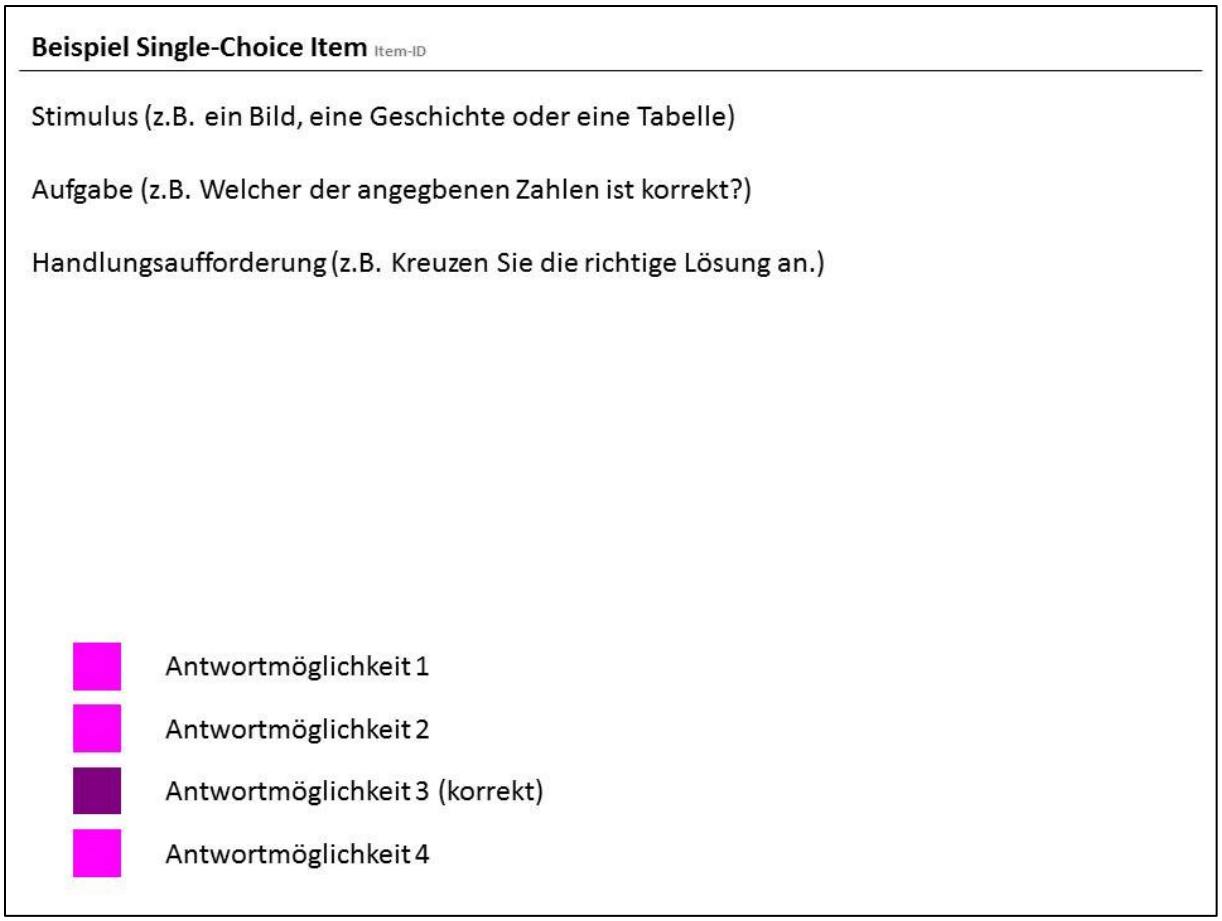

Abbildung 6. Vorlage für das Layout eines Items in Microsoft PowerPoint.

Bei der Erstellung der Items für MATE durch die Software Microsoft PowerPoint muss auf gewisse Punkte geachtet werden. In Bezug auf die Antwortmöglichkeiten eines Items erzeugen Vierecke in der Farbe Magenta grafische Optionsfelder (Radiobuttons). Antwortmöglichkeiten, welche mit dem dunklen Magenta markiert wurden, werden als korrekte Antwort hinterlegt. Die Folien werden stets im Querformat gesetzt. Der Titel ist hier eine Kurzüberschrift für das Item. Direkt daneben wurde in kleinerer grauer Schrift die interne Item-ID abgebildet. Der Stimulus enthält eine Frage, ein Statement, eine Abbildung oder eine Tabelle, in der Informationen zum Lösen der Aufgabe gegeben werden. Die Schriftgröße wurde mit 18 Punkten ausreichend groß gewählt. Schriftgrößen in Abbildungen und Tabellen können auch eine Schriftgröße kleiner als 18 Punkte aufweisen. Es sollte aber immer darauf geachtet werden, dass die Inhalte auch auf kleineren Monitoren problemlos lesbar sind. Die Iteminhalte sollten möglichst auf einer Seite bzw. Folie untergebracht werden, damit ein Vor- und Zurückblättern zwischen den Seiten nicht notwendig ist. Nach der Erstellung aller Items aus der Itemdatenbank in Microsoft PowerPoint wurden die Items in das Dokumentenformat XPS umgewandelt.

Dieses Format kann problemlos von der Software MATE eingelesen werden. Nach diesem Schritt sind die Items als grafischer Hintergrund einlesbar. Alle zusätzlichen Informationen zu Itemnamen, korrekte Lösung des Items bei offenen Items, Positionen, Funktionen und Beschriftung von Navigations-Buttons, Reihenfolge der Items, Inhaltsbereich für Content-Balancing, Zulassen von Itemreview, Itemparameter usw. müssen über eine sogenannte Schlüsseldatei eingelesen werden. Diese Datei liegt zusammen mit den XPS-Dateien (Items) in einem Ordner und kann dann über Software MATE importiert werden. Nach dem Import der Items und der Schlüsseldatei in die Software MATE sah ein Item wie folgt aus:

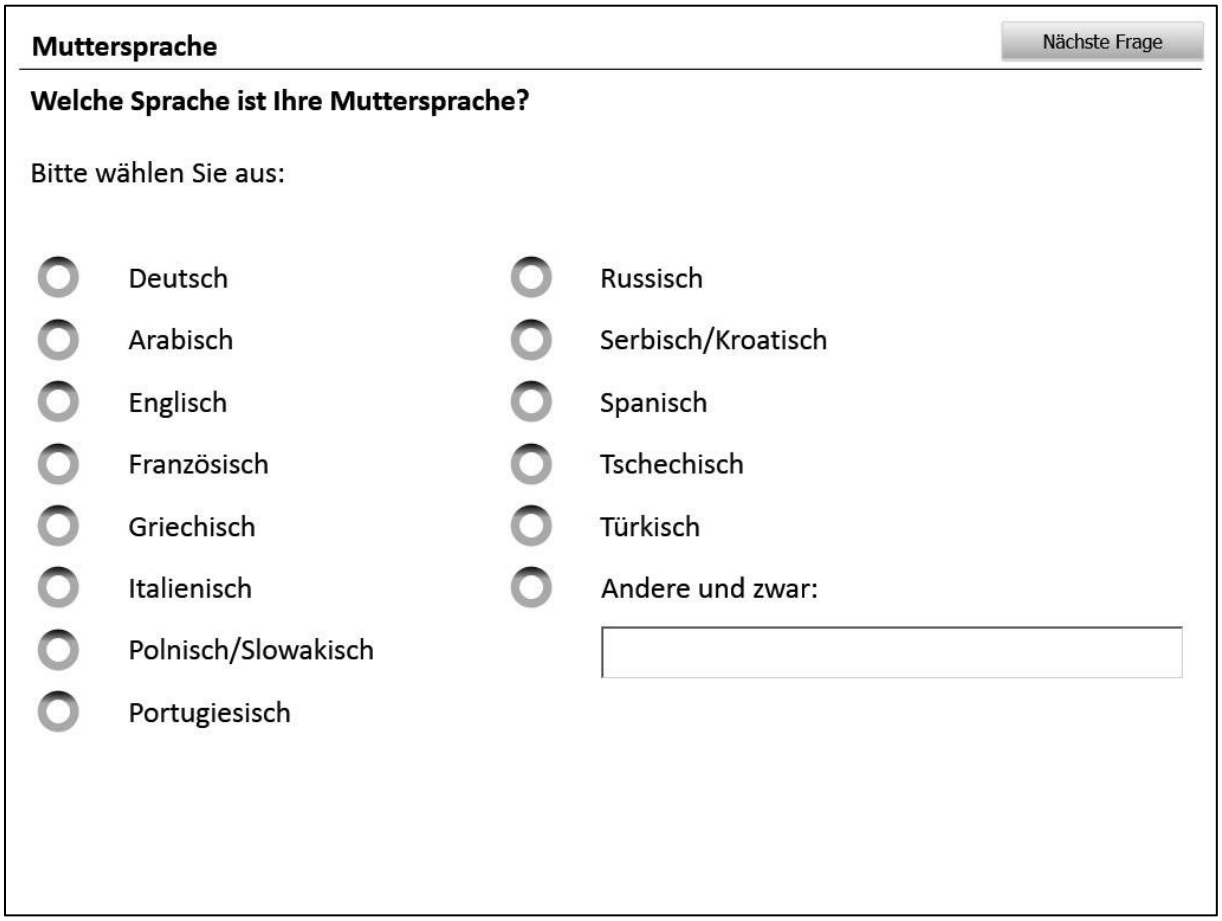

Abbildung 7. Beispielitem nach dem Einlesen in MATE.

Als Beispielitem wurde ein Item zur Abfrage der *Muttersprache* abgebildet. Im Item Muttersprache sind die beiden verwendeten Itemformate bei MaK-adapt zu erkennen (Single-Choice und Einfache offene Antwortformate). Kompetenzitems wurden aufgrund der Testsicherheit nicht abgebildet.

4.2.4 Zusammenfassung

In diesem Abschnitt wurde verdeutlicht, wie mit wenigen Ressourcen in kurzer Zeit ein Itempool für die drei Domänen Lesen, Mathematik und Naturwissenschaft von insgesamt mehr als 300 Items entwickelt werden konnte. Itemrecycling war hier die Methode der Wahl für einen Großteil der Items. Mit der Anpassung vorhandener Items aus anderen Studien konnten mehrere Monate Zeit und viele Entwicklungskosten gespart werden. Ein weiterer Vorteil des Itemrecycling liegt darin, dass die meisten verwendeten Items bereits in anderen Studien einen Pretest überstanden haben und so die Quote der Items, die anschließend noch ausgeschlossen werden müssen, geringer ausfällt. So konnte der geplante Zeitrahmen eingehalten und trotzdem viele Items generiert werden. Bei der Zusammenstellung der Items für den Itempool wurden zuvor die Bearbeitungszeiten und die Schwierigkeit der Items eingeschätzt. Die Schätzung von Bearbeitungszeiten und Schwierigkeiten der Items bringt etwas mehr Arbeit mit sich. Doch wegen des engen Zeitplans, in dem nur eine Kalibrierung inklusive Pretest vor der Pilotierung vorgesehen war, konnten nicht erst die empirischen Ergebnisse abgewartet werden. Eine Gleichverteilung der Items über die Itemschwierigkeiten wurde angestrebt. Ein Hinweis darauf, dass nach der Kalibrierungsstudie Items in einem gewissen Schwierigkeitsbereich fehlen oder viele zu lange Items vorhanden sind, wäre anschließend zeitlich nicht mehr auszugleichen gewesen.

Als Antwortformat wurden überwiegend Single-Choice-Items verwendet und wenige offene Items. Offene Items wurden jedoch nur bei Antworten gewählt, wo eine einfache Antwort (z. B. die Zahl 8) richtig war und als richtige Antwort möglichst wenig Alternativen in der MATE hinterlegt werden mussten (z. B. 8, acht, Acht und weitere falsch geschriebene Möglichkeiten, die aus der Kalibrierungsstudie hervorgingen). Inhaltlich wurde der Stimulus aus motivationalen Aspekten häufig durch Bilder und Tabellen ansprechend gestaltet. Medieninhalte, wie Videos oder Audiodateien, konnten in der Software MATE aus technischen Gründen nicht hinterlegt werden. Die Computerisierung der Items konnte aufgrund der komfortablen Möglichkeiten von MATE einfach und schnell durchgeführt werden. Mit Hilfe von Microsoft PowerPoint konnten die Items schnell gesetzt werden und ein langes Einarbeiten in eine spezielle Software war nicht notwendig. Die Verwendung eines Layouts und einer Schlüsseldatei führte dazu, dass die Setzung der unterschiedlichen Inhalte der Items (Stimulus, Antwortformate, Buttons

usw.) ohne viel Aufwand und Zeit stets gleich waren. Durch die Erstellung einer Schlüsseldatei konnten die Items zusammen in kurzer Zeit in die MATE importiert werden. Für das Itempoolmanagement wurde die Software Microsoft Excel verwendet und ein ausführliches Item-ID-System hier vorgestellt. Diese beiden Werkzeuge vereinfachen das Arbeiten mit Items über die Zeit und innerhalb mehrerer Projekte erheblich. Es ist darauf zu achten, dass die Item-IDs verständlich formuliert und konsistent verwendet werden und jede ID nur einmal benutzt wird. Die ID sollte so geplant sein, dass auch in Zukunft noch Items hinzukommen können, ohne die Struktur der ID ändern zu müssen. Gleiches gilt für die Datenbank, in welcher der Itempool gespeichert wird. Es wird empfohlen die Items in einer Datenbank so zu managen, dass über die Zeit hinweg alle wichtigen Informationen dort hinterlegt sind und die Datenbank stetig erweitert werden kann. Ein Vorschlag hierzu wurde angeführt.

Ziel bei der Entwicklung des Itempools war es, neben qualitativ hochwertigen Items, welche das inhaltliche Zielkonstrukt adäquat messen, auch genügend Items in den entsprechenden Schwierigkeitskategorien zu haben. Es wurde beispielhaft gezeigt, wie man die notwendige Anzahl an Items ermitteln kann. Jedoch ist diese Frage aufgrund von unterschiedlichen Bedienungen nicht allgemeine zu beantworten. Im Projekt MaK-adapt wurde in den Domänen Mathematik und Naturwissenschaft angestrebt, ca. 20 Items in jedem theoretischen Schwierigkeitsbereichen von sehr leicht, leicht, durchschnittlich, schwer und sehr schwer und ca. 25 Items in jeder der vier Subdomäne zu haben. Zudem wurden bei einem angestrebten Itempool von 100 Items ca. 30 % mehr Items in die Kalibrierungsstudie eingebracht, da häufig nach dem Pretest und der Kalibrierung noch Items aus dem Pool entfernt werden. Bei Lesen wurden deutlich weniger Items eingebracht, was zum einem der geringeren Anzahl an Subdomänen und zum anderen dem aufwendigeren Prozess der Itementwicklung in dieser Domäne geschuldet ist. Außerdem ist bei den Leseitems ein deutlich höherer Lese- und somit Bearbeitungsaufwand zu erwarten, weshalb bei der Kalibrierung und im späteren computerisierten adaptiven Test bei gleicher Testzeit weniger Items vorgelegt werden können.

4.3 Pretest und Kalibrierung des Itempools

Nach der Entwicklung des initialen Itempools, können die Items getestet und kalibriert werden. Kalibrieren bedeutet in diesem Zusammenhang, die Itemparameter festlegen. Denn beim adaptiven Testen bekommen Testpersonen Items vorgelegt, die ihrem Kompetenzstand bestmöglich entsprechen. Die Itemschwierigkeit ist somit dem Kompetenzniveau angepasst (vgl. Kapitel 3.1.2). Dafür werden bereits vor der Testung Itemparameter benötigt. Diese müssen neben den Items auch in der verwendeten Software hinterlegt werden. Um die Items im Feld zu testen und die benötigten Itemparameter empirisch zu schätzen, wurde im Projekt MaK-adapt eine Kalibrierungsstudie durchgeführt. Ziel der Kalibrierungsstudie war es, die Itemparameter möglichst präzise für die untersuchte Population zu schätzen sowie defizitäre Items zu identifizieren und aus dem Itempool zu entfernen. Aufgrund der hohen Anzahl an zu kalibrierenden Items können nicht alle vorhandenen Items in angemessener Zeit durch jeden Probanden bearbeitet werden. Deshalb wurde ein Testheftdesign entwickelt, so dass alle Items gleich häufig an jeder Position vorgegeben werden können, aber jeder Proband der Kalibrierungsstudie lediglich maximal 33 Items beantworten muss. In diesem Abschnitt werden das Testheftdesign, die Stichprobe, die Methode und die Ergebnisse der Kalibrierungsstudie vorgestellt. Zusätzlich wird hier ein weiterführender Schritt, die Analyse von Positionseffekten bei der Entwicklung eines computerisierten adaptiven Tests, eingeführt und beispielhaft dargestellt.

4.3.1 Fragestellungen

- Wie ist das Testheftdesign bei einer großen Anzahl an zu kalibrierenden Items zu wählen, wenn die Möglichkeit bestehen soll, Positionseffekte auf Itemebene zu modellieren?
- Wie sind die Probanden der Kalibrierungsstudie verteilt?
- Welches IRT-Modell eignet sich zur Kalibrierung der vorliegenden Daten?
- Was ist bei der Itemselektion zu berücksichtigen?
- Wie lässt sich DIF identifizieren?
- Wie sind die Items im Itempool hinsichtlich Inhalt und Schwierigkeit nach der Kalibrierung verteilt?

- Liegen Itempositionseffekte vor?
- Sind vorliegende Itempositionseffekte für alle Items gleich?
- Wie wirkt sich die Modellierung von Positionseffekten auf die Skalen (Varianz, Reliabilität) aus?
- Wie groß sind die Itempositionseffekte?
- Hat die Betrachtung von Itempositionseffekten Auswirkung auf die Itemkennwerte?

4.3.2 Testheftdesign

In der Kalibrierungsstudie wurde ein unvollständiges balanciertes Testheftdesign mit zwei Ebenen genutzt. Ziel war es, alle Items auf allen möglichen Positionen im Testheft gleichmäßig häufig vorzugeben. Zusätzlich sollte auch eine multidimensionale Schätzung möglich sein. Deshalb wurde ein Testheftdesign mit zwei Ebenen verwendet. Auf der ersten Ebene wurden die drei Domänen Lesen, Mathematik und Naturwissenschaft vollständig permutiert (vgl. Tabelle 4).

Tabelle 4

Testheftdesign auf Ebene 1 (L- Lesen, M- Mathematik, N- Naturwissenschaft)

	Testheft Ebene 1					
Position im Testheft	1	2	3	4	5	6
1	L	L	N	N	M	M
2	M	N	L	M	N	L
3	N	M	M	L	L	N

Durch dieses Vorgehen werden unidimensionale Testblöcke verwendet und es ist zugleich eine multidimensionale Schätzung möglich. Es gibt somit sechs Testhefte auf Ebene 1. Dabei enthält jedes Testheft alle drei Domänen an unterschiedlichen Positionen. Auf der zweiten Ebene wurden die Items mit einem Youden-Square-Design (YSD; vgl. Kapitel 3.4.1) innerhalb der Domänen gleichmäßig verteilt. Die Parameter des YSD waren für die Domäne Lesen $t = b = 73,\ k = r = 9,\ \lambda = 1$ und für die Domänen Mathematik und Naturwissenschaft $t = b = 133,\ k = r = 12,\ \lambda = 1$. Konkret heißt das z.B. für die Domäne Lesen, dass jedes der insgesamt 73 Items t genau einmal in 73

unterschiedlichen Testheften b auftaucht. Jedes Item erscheint $r = 9$ mal über alle Testhefte hinweg und jedes Testheft hat eine Länge von $k = 9$ Items. Jedes Paar von Items (λ) taucht in den Testheften maximal einmal auf. Das vorgestellte Design auf zwei Ebenen ermöglicht sowohl die gleichmäßige Verteilung der Items auf Positionsebene im Testheft als auch die die Balancierung der Items auf Itemebene. Insgesamt gab nach der Schachtelung der Testhefte aus den beiden Ebenen 798 verschiedene Testhefte mit jeweils 33 Items (12 Mathematikitems, 12 Naturwissenschaftsitems und neun Leseitems).

4.3.3 Ablauf und Stichprobe: Kalibrierungsstudie

Die Testung der SuS an den beruflichen Schulen erfolgte onlinebasiert über einen Browser am Computer in den Computerräumen der entsprechenden Schulen. Der Test wurde über eine URL aufgerufen, heruntergeladen und vor der Testung auf den entsprechenden Computern gespeichert. Die Antworten der Probanden wurden nach jeder Eingabe an einen Server in Jena gesendet. Bei einer Unterbrechung der Verbindung wurden die Daten solange lokal zwischengespeichert, bis die Verbindung wieder hergestellt wurde. Da für die Testungen die Computertechnik und die Internetverbindung der entsprechenden Schulen verwendet wurden, musste eine technische Überprüfung vor der eigentlichen Testung durchgeführt werden, um einen reibungslosen Ablauf am Testtag gewährleisten zu können (vgl. Kapitel 4.1.3). Je nach Ausstattung der Schulen kam es zu technischen Problemen, welche teilweise zuvor gelöst werden konnten. Probleme waren u. a.

- eine schlechte Qualität der Internetverbindung, so dass nicht für alle SuS im Computerraum gleichzeitig die Testhefte auf den Computer geladen werden konnten,
- fehlende Administratorrechte, so dass die notwendige Software für die Testdurchführung nicht auf den Computern installiert werden konnte oder
- restriktive Firewall-Einstellungen, welche teilweise das Senden und Empfangen der Daten verhinderten.

Schulen, die aufgrund der technischen Überprüfung an der Testung teilnehmen konnten, wurden am Testtag von mindestens zwei Testleitern pro Computerraum unterstützt. Die Testleiter waren dafür zuständig, Fragen während der Testung zu beantworten, bei

Problemen zu helfen und eine angemessene Testatmosphäre herzustellen. Für die Kalibrierungsstudie wurden eine Teilnehmerzahl von $N \geq 1\,000$ SuS angestrebt. Die Ziehung der Teilnehmer erfolgte nach einem Stichprobenplan. Es wurden in erster Linie Schulen gewählt, die SuS in den anvisierten ASCOT-Berufen ausbilden (Kfz-Mechatroniker/in, Elektroniker/in für Automatisierungstechnik, Industriekaufmann/-frau, Pflegekräfte für ältere Menschen und Medizinische Fachangestellte). Zusätzlich wurden Schulen gewählt, die ähnliche Berufe ausbilden, um die Stichprobengröße zu erhöhen. Ein weiteres Kriterium war, dass die Schulen in Niedersachsen, Hessen und Thüringen lagen und dass die Schulen in einer Klasse mindestens 20 SuS in den ausgewählten Berufen haben, damit sich die Anfahrt rentiert. Vor allem SuS im letzten Ausbildungsjahr ihres Ausbildungsganges sollten in die Stichprobe eingehen. Die Ausbildung dauert bei den meisten Ausbildungsberufen drei Ausbildungsjahre, an manchen Schulen bzw. in machen Berufen sind zwei oder vier Ausbildungsjahre vorgesehen.

Bei der Kalibrierungsstudie bekamen $N = 1\,632$ Personen an 27 berufsbildenden Schulen in den Bundesländern Niedersachsen, Hessen und Thüringen einen computerisierten adaptiven Test entsprechend dem Testheftdesign vorgelegt. Die Testhefte mit 33 Items wurden durchschnittlich in 21 Minuten bearbeitet ($SD = 9$ Minuten). Für die Domäne Lesen wurde eine durchschnittliche Bearbeitungszeit von 97 sek pro Item ermittelt. Für die Mathematikitems wird mit einer Bearbeitungszeit von durchschnittlich 62 sek pro Item und für die Naturwissenschaftsitems mit durchschnittlich 43 sek pro Item deutlich weniger Zeit benötigt. Das Durchschnittsalter der getesteten SuS beträgt 21.384 Jahre ($SD = 3.032$ Jahre). Die weiteren Häufigkeitsangaben zur Beschreibung der Stichprobe sind zur besseren Lesbarkeit als Stichpunkte dargestellt:

- Ausbildungsjahr: 6.7 % viertes Ausbildungsjahr; 66.3 % drittes Ausbildungsjahr; 20.7 % zweites Ausbildungsjahr; 5.1 % erstes Ausbildungsjahr; 1.2 % keine Angabe
- Geschlecht: 46.3 % weiblich; 52.6 % männlich; 1.2 % keine Angabe
- Schulabschluss: 28.5 % allgemeine Hochschulreife bzw. Fachhochschulreife; 62.1 % mittlere Reife; 7.2 % Haupt- bzw. Volksschulabschluss; 0.5 % ohne Schulabschluss oder Abschluss der Sonderschule bzw. Förderschule; 1.7 % keine Angabe
- Muttersprache: 86.9 % Deutsch; 11.3 % andere Sprache; 1.8 % keine Angabe

- Form der Berufsausbildung: 94.4 % duale Berufsausbildung; 4.4 % vollzeitschulische Berufsausbildung; 1.2 % keine Angabe
- Anzahl der Beschäftigten im Ausbildungsbetrieb: 19.2 % weniger als 10 Beschäftigte; 23.3 % zwischen 10 und 49 Beschäftigte; 22.1 % zwischen 50 und 249 Beschäftigte; 8.5 % zwischen 250 und 499 Beschäftigte; 21.6 % mit 500 und mehr Beschäftigten; 5.2 % keine Angabe oder in vollzeitschulischer Berufsausbildung
- Standort des Ausbildungsbetriebs: 24.4 % Hessen; 42.3 % Niedersachsen; 30.0 % Thüringen; 2.2 % anderes Bundesland; 1.1 % keine Angabe
- Berufsfeld: 22.9 % medizinisch/pflegender Bereich; 38.2 % gewerblich/technischer Bereich; 33.6 % kaufmännisch/verwaltender Bereich; 4.0 % anderes Berufsfeld; 1,3 % keine (plausible) Angabe
- Innerbetrieblicher Unterricht: 56.0 % innerbetrieblicher Unterricht; 42.7 % kein innerbetrieblicher Unterricht; 1. 3 % keine Angabe

4.3.4 Methode und Ergebnisse: Kalibrierungsstudie

Nach der Durchführung der Kalibrierungsstudie und der Speicherung der Daten auf einem Server wurden die Daten mit Hilfe der Software SPSS für weitere Analysen aufbereitet. Bei der Behandlung der fehlenden Werte wurde die Bearbeitungszeit der Items mit berücksichtigt. Es wurde angenommen, dass ein Proband, der ein Item bearbeitet, eine gewisse Zeit benötigt, um das Item zu sichten und sich Gedanken zu der Antwort zu machen. Nach Durchsicht der Items wurde für den hier verwendeten Itempool angenommen, dass ein Proband durchschnittlich mindestens fünf Sekunden für die Betrachtung eines Items benötigt, um eine verlässliche Einschätzung über die Beantwortung des Items treffen zu können. Auf Grundlage dieser Theorie wurde festgelegt, dass fehlende Antworten auf Items, die kürzer als fünf Sekunden angesehen wurden, als fehlende Werte behandelt werden. Wurde ein Item mit fehlender Antwort fünf Sekunden oder länger angeschaut, wurde angenommen, dass es theoretisch bearbeitet werden konnte. Deshalb wurde in diesen Fällen der fehlende Wert als falscher Wert umcodiert. D. h., es wurde dem Probanden unterstellt, dass er sich das Item angeschaut hat und bewusst im Test weitergegangen ist, ohne eine Antwort zu geben.

Nach der Behandlung der fehlenden Werte und der Datenaufbereitung wurden die Daten für die Weiterverarbeitung in der Software ConQuest vorbereitet. Mit Hilfe von ConQuest konnten unterschiedliche Modelle der IRT mit den Daten gerechnet und verglichen werden. Hier wurde sich für ein eindimensionales Rasch-Modell (vgl. Formel (1) auf S. 18) zur Skalierung der Daten entschieden. In ConQuest sieht der Ausschnitt aus der Syntax zur Ermittlung der Itemparameter wie folgt aus:

```
set constraints=cases;

model item;
```

Zur Ausführung der Schätzung in ConQuest sind weitere Befehle notwendig, welche dem ConQuest-Manual (Wu et al., 2007) entnommen werden können. Es ist notwendig, die Einschränkung (set constraints) auf die Personen (cases) zu beziehen. Die Einschränkung bewirkt, dass der letzte Personenparameter fixiert und so der Mittelwert der Personenparameter 0 wird. Auf diese Weise können alle Itemparameter frei geschätzt werden. Bei der Festlegung der Einschränkung auf die Items würde der Parameter des letzten Items nicht frei geschätzt werden. Für die Kalibrierung der Items ist es jedoch von Bedeutung, dass alle Items frei geschätzte Itemparameter erhalten, die später im adaptiven Algorithmus ohne Einschränkungen verwendet werden können. Nach der Skalierung der Items wurden die Ergebnisse aus der Kalibrierungsstudie dazu verwendet, anhand verschiedener statistischer und inhaltlicher Kriterien, defizitäre Items zu identifizieren und aus dem Itempool zu entfernen. Konkret wurden Items mit geringer Diskrimination/ Trennschärfe ($item\ total\ correlation\ r < .25$), einer hohen punktbiserialen Korrelation ($r_{pb} > .1$) mit einem Distraktor (falsche Antwortkategorie) und einem schlechtem Itemfit ($t_{WMNSQ} > 1.96$) identifiziert und ggf. entfernt. Der Itemfit wurde lediglich auf WMNSQ-Werte, die signifikant höher als der Wert 1 sind, geprüft. In diesem Fall ist die tatsächliche ICC aus den empirischen Daten flacher als die erwartete ICC. Items mit hohen WMNSQ-Werten wurden anschließend inhaltlich geprüft und bei sichtbaren Fehlern (z. B. zwei richtige Antwortkategorien im Item) entfernt. Items mit hohem WMNSQ und ohne sichtbaren Fehler wurden nur dann entfernt, wenn sie aus einem Schwierigkeitsbereich oder einer Subdomäne stammen, in der bereits ausreichend viele Items vorhanden waren. Items mit niedrigen WMNSQ-Werten wurden aus pragmatischer Sicht im Itempool gelassen (vgl. Kapitel 3.4.2). Ziele dieses Vorgehens

waren eine gleichmäßige Abdeckung der theoretischen Kompetenzmodelle und eine gleichmäßige Abdeckung eines möglichst breiten Schwierigkeitsspektrums. Im Zuge der Itemselektion wurden die Items im nächsten Schritt auf DIF untersucht. Dies geschah zuerst im Hinblick auf das Geschlecht. Dazu wurde ein Multifacetten-Rasch-Modell mit der zusätzlichen Facette Geschlecht (gender) mit den Ausprägungen männlich und weiblich betrachtet (vgl. Formel (11) auf S. 53). In ConQuest wird die Zeile zur Spezifikation des Modells folgendermaßen abgeändert:

```
model item + gender + item*gender;
```

Der Ausdruck *item*gender* entspricht dabei dem Interaktionseffekt $G_g b_i$ aus Formel (11) auf S. 53, welcher den Effekt der mittleren Fähigkeit G_g der Gruppe g und der Itemschwierigkeit b_i für das Item i wiedergibt. In Bezug auf DIF drückt dieser Wert aus, wie unterschiedlich die Wahrscheinlichkeit ausfällt, ein Item korrekt zu beantworten, nachdem die mittleren Kompetenzunterschiede zwischen den Gruppen männlich und weiblich berücksichtigt wurden. Im Idealfall sollte es nach Berücksichtigung der mittleren Kompetenzunterschiede keine Unterschiede mehr in der Wahrscheinlichkeit geben ($G_g b_i = 0$). Eine signifikante Abweichung von dem Wert 0 kann somit als ein Hinweis auf DIF gewertet werden. Die Abweichung von 0 wurde auf einem Signifikanzniveau von $\alpha = .01$ geprüft. In weiteren Modellen zur Geschlechter-DIF-Analyse wurden die weiteren Haupteffekte Beruf (job; Modell A) sowie Beruf und Muttersprache (job + language; Modell B) hinzugefügt. So konnte geprüft werden, ob ein DIF-Effekt in Bezug auf die Variable Geschlecht vorliegt, nachdem die Variablen Beruf und Muttersprache als Haupteffekte herausgerechnet wurden. In der Syntax in ConQuest wurde dazu jeweils die Zeile zum Modell angepasst:

Modell A: `model item + gender + job + item*gender;`

Modell B: `model item + gender + job + language + item*gender;`

Alle statistisch identifizierten Items wurden anschließend inhaltlich geprüft. Die inhaltliche Prüfung erfolgte, indem eine gerichtete Hypothese aufgestellt wurde, dass die Itemschwierigkeit geschlechtsspezifisch ist. Beispielsweise kann man bei einem Item zum Thema Fußball die Hypothese aufstellen, dass das Item für Männer einfacher ist. Anschließend erfolgte die Prüfung der Hypothese über den Interaktionseffekt $G_g b_i$

(item*gender). Zeigte der Interaktionseffekt dieselbe Richtung wie die zuvor gebildete Hypothese, wurde das Item aus dem Itempool entfernt. Eine DIF-Analyse für die Variable Muttersprache in einer gleichen Weise durchzuführen wie mit der Variable Geschlecht erwies sich aus statistischer Sicht als schwierig, da 86.9 % der SuS angaben, als Muttersprache Deutsch zu sprechen. Es gab somit Items mit wenigen bis gar keinen Antworten von Personen mit nicht deutscher Muttersprache. Die DIF-Analyse für die verschiedenen Gruppen von Ausbildungsberufen erfolgte in einer weiteren ausführlichen Studie ebenfalls mit einem Multifacetten-Rasch-Modell. Dabei wurde die Variable Ausbildungsberuf dichotom in technisch-gewerblich und kaufmännisch-verwaltend unterteilt. Zusätzlich wurde eine Kontrollvariable mit aufgenommen, welche die Position eines Items im Testheftdesign auf Domänenebene (Testheftebene 1; vgl. Tabelle 4 auf S. 108) angibt. Die statistisch identifizierten Items wurden anschließend durch Inhaltsexpertinnen auf DIF geprüft. Das Vorgehen und die Ergebnisse hierzu finden sich ausführlich beschrieben bei Spoden et al. (2015). Anzumerken ist, dass bei der DIF-Analyse zusätzlich zur statistischen Identifikation immer auch eine inhaltliche Analyse erfolgen sollte, bevor Items aus dem Itempool entfernt werden. Die Analyse des WMNSQ auf einen signifikant höheren Wert als 1 ergab, dass ein Leseitem, zwei Mathematikitems und kein Naturwissenschaftsitem aufgrund fehlender Passung zum Rasch-Modell identifiziert wurden. Im Rahmen der DIF-Analysen wurden sechs Leseitems, neun Mathematikitems und neun Naturwissenschaftsitems identifiziert. Ausschließlich aus DIF-Gründen wurden bei Lesen ein Item, bei Mathematik zwei Items und bei Naturwissenschaft drei Items entfernt. Aufgrund von zu geringer Trennschärfe ($r < .25$) wurden kein Item in Lesen, 11 Items in Mathematik und 16 Items in Naturwissenschaft identifiziert. Bei drei Mathematikitems war aufgrund der niedrigen Schwierigkeit keine hohe Diskrimination zu erwarten. Deshalb wurden letztendlich nur acht Mathematikitems aufgrund zu geringer Trennschärfe identifiziert. Aufgrund zu hoher punktbiserialen Korrelation ($r_{pb} > 0.1$) mit einem Distraktor wurden zwei Leseitems (zwei Items entfernt), 15 Mathematikitems (sieben Items entfernt) und 18 Naturwissenschaftsitems (fünf Items entfernt) identifiziert. Einige der Items wurden aufgrund mehrerer statistischer Kriterien identifiziert und andere Items ohne Identifikation aufgrund zusätzlicher inhaltlicher Überlegungen entfernt. Deshalb muss die Summe der identifizierten Items nicht der Anzahl der tatsächlich entfernten Items entsprechen. Nach der Itemselektion aufgrund der oben beschriebenen Kriterien, der DIF-Analyse und inhaltlicher Überle-

gungen blieben für die Domäne Lesen 93.2 % ($N = 68$), für die Domäne Mathematik 85.0 % ($N = 113$) und für die Domäne Naturwissenschaft 73.3 % ($N = 96$) der ursprünglichen Items im Itempool. In der Domäne Lesen waren vor allem in der Subdomäne *Gemischte Darbietung* viele Items vorhanden.

Tabelle 5

Verteilung der Items nach Itemselektion für die Domäne Lesen

Subdomäne	Anzahl ausgewählter Items
Deskriptional	23
Gemischte Darbietung	26
Depiktional	19

Für die Domäne Mathematik war es schwierig, genügend Items für die Subdomäne *Unsicherheit* zu finden. Es lagen zwar Items aus anderen Studien vor. Jedoch besaßen diese überwiegend ein komplexes offenes Antwortformat, welches nicht automatisiert ausgewertet werden konnte.

Tabelle 6

Verteilung der Items nach Itemselektion für die Domäne Mathematik

Subdomäne	Anzahl ausgewählter Items
Quantität	30
Veränderung & Beziehung	29
Raum & Form	30
Unsicherheit	24

Bei der Domäne Naturwissenschaft waren mit 20 Items in der Subdomäne *Bewegung, Energie, Kraft* die wenigsten Items im Pool vorhanden.

Tabelle 7
Verteilung der Items nach Itemselektion für die Domäne Naturwissenschaft

Subdomäne	Anzahl ausgewählter Items
Leben, Gesundheit	26
Erde, Planeten, Umwelt, natürliche Ressourcen	22
Stoffe, Stoffveränderungen	28
Bewegung, Energie, Kraft	20

Insgesamt betrachtet ist die erzielte Verteilung der Items über die Subdomänen dennoch ein Erfolg hinsichtlich der Gleichverteilung der Items, da jede Subdomäne eine ausreichende Substanz an Items besitzt, um später den MPI als Content-Balancing-Methode anwenden zu können. Die Verteilung des Schwierigkeitsparameters b sollte entsprechend der Theorie über den gewählten Schwierigkeitsbereich (hier -5 Logits bis 5 Logits) möglichst gleich verteilt sein, damit der adaptive Algorithmus in jedem Fähigkeitsbereich genügend Items zur Auswahl hat (Frey, 2012). Für die Domäne Mathematik lag der Mittelwert der Itemschwierigkeit bei $Mean(b_{MATH}) = -0.278$ und die Standardabweichung bei $SD(b_{MATH}) = 1.324$. Für die Domäne Naturwissenschaft betrugen der Mittelwert $Mean(b_{SCIE}) = -0.720$ und die Standardabweichung $SD(b_{SCIE}) = 1.102$. In der Domäne Lesen lag die mittlere Itemschwierigkeit bei $Mean(b_{READ}) = -0.378$ und die Standardabweichung bei $SD(b_{READ}) = 1.119$. In den nachfolgenden Abbildungen ist die relative Häufigkeit der Items pro Itemschwierigkeitsbereich in Logits sowie die relative Häufigkeit der Personen pro Fähigkeitsbereich (WLE) in Logits zu sehen. Dabei wurden Items mit dicht beieinanderliegenden Schwierigkeitsparametern in 0.5 Logit-Schritten zusammengefasst. Für die Domäne Lesen ist zu sehen, dass in den Randbereichen der Verteilung der Itemschwierigkeit Items fehlen. In Zusammenhang mit den WLEs bildet der Itempool die Zielgruppe jedoch relativ gut ab. Zum Auffüllen des Itempools sollten vor allem Items aus dem oberen Schwierigkeitsbereich (zwischen einem und vier Logits) erstellt werden.

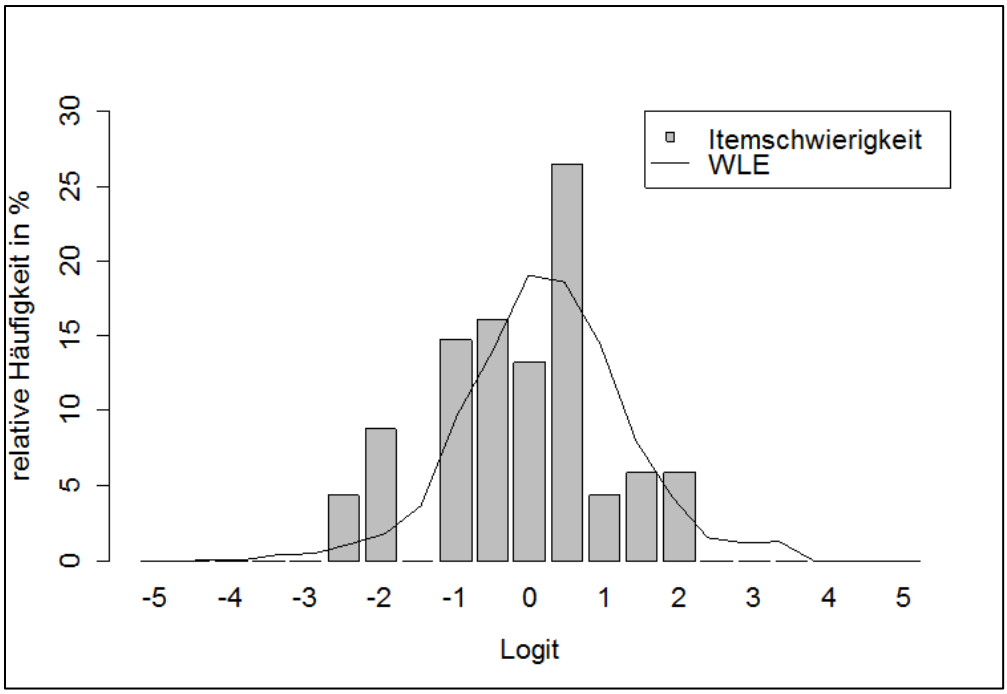

Abbildung 8. Prozentuale Anzahl an Items pro Schwierigkeitsparameter für die Domäne Lesen.

In der Domäne Mathematik sind beinah in allen Schwierigkeitsbereichen Items vorhanden. Die Mehrzahl der Items verteilt sich auf den mittleren Schwierigkeitsbereich.

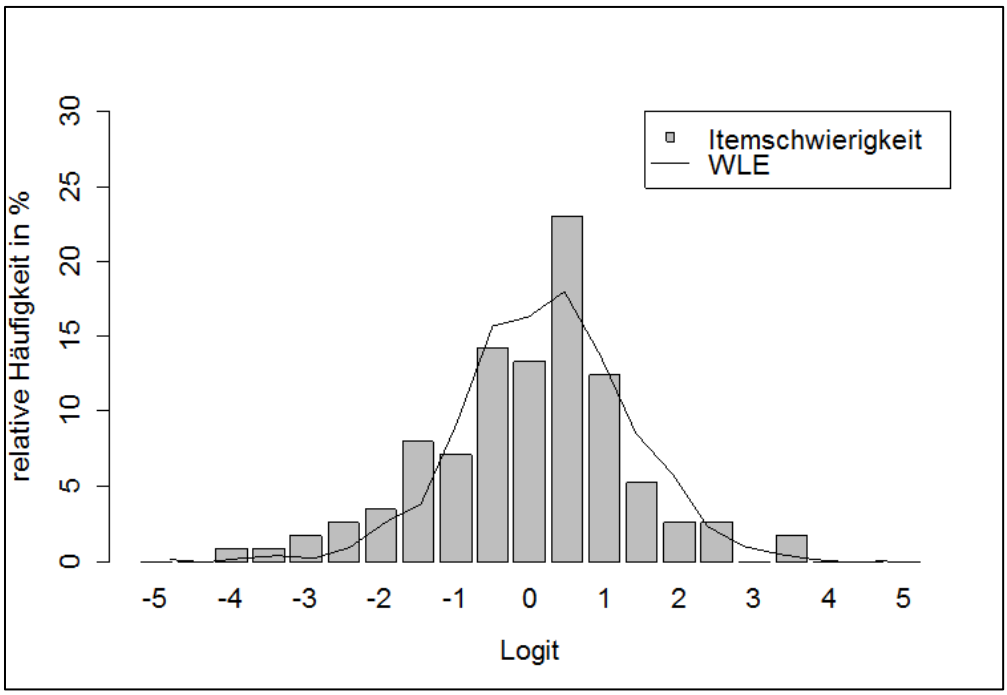

Abbildung 9. Prozentuale Anzahl an Items pro Schwierigkeitsparameter sowie relative Häufigkeit der WLE für die Domäne Mathematik.

Die Domäne Naturwissenschaft enthält Items im Schwierigkeitsbereich zwischen -2,5 und 2. Dort sind die Items relativ gleich verteilt. Im Randbereich fehlen Items. Vor allem sind in der untersuchten Stichprobe vermehrt Personen im oberen Fähigkeitsbereich über 0 Logits und im Vergleich dazu relativ wenig Items vorhanden.

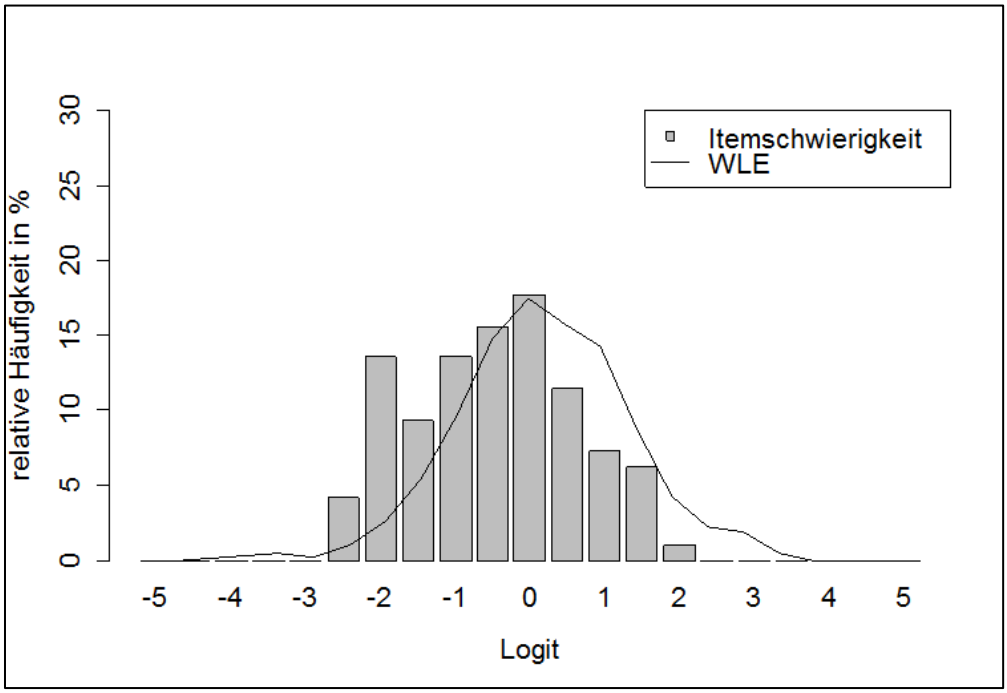

Abbildung 10. Prozentuale Anzahl an Items pro Schwierigkeitsparameter für die Domäne Naturwissenschaft.

Nach der Pilotierungsstudie (vgl. Kapitel 4.5) wurden im Projekt MaK-adapt auf Basis eindimensionaler Skalierungen mit dem Rasch-Modell durch einen zusätzlichen Selektionsprozess weitere defizitäre Items (ein Leseitem und ein Naturwissenschaftsitem) identifiziert und aus dem Itempool ausgeschlossen. Zudem wurden die Items mit einem komplexen Multiple-Choice Antwortformat aufgrund von technischen Hürden in der Software MATE zu der Zeit der Pilotierungsstudie ausgeschlossen (Mathematik acht Items; Lesen zwei Items; Naturwissenschaft ein Item). Im Ergebnis wurden 65 Leseitems, 105 Mathematikitems und 94 Naturwissenschaftsitems in den adaptiven Tests im ASCOT-Verbund verwendet. Ein Item im Itempool enthält durchschnittlich in der Domäne Lesen 208.738 Wörter (σ=123.218) und 2.185 Seiten (σ=0.926), in der Domäne Mathematik 60.591 Wörter (σ=30.005) und 1.124 Seiten (σ=0.329) und in der Domäne Naturwissenschaft 60.553 Wörter (σ=39.668) und 1.106 Seiten (σ=0.308). Ein oder mehrere Bilder waren in 53.846 % der Leseitems, 51.429 % der Matheitems und

28.723 % der Naturwissenschaftsitems enthalten. In der Domäne Lesen gibt es 53 Single-Choice-Items und 12 offene Items. In der Domäne Mathematik haben 13 von 105 Items eine kurze offene Antwort, die restlichen 92 Items besitzen den Antwortmodus Single-Choice (bzw. einfaches Multiple Choice). In der Domäne Naturwissenschaft gibt es 93 Single-Choice-Items und ein offenes Item. Die nachfolgenden Ergebnisse zu den Positionseffekten beziehen sich auf den genannten reduzierten Itempool, wie er in der ASCOT-Initiative verwendet wurde.

4.3.5 Methode und Ergebnisse: Positionseffekte

Da das Vorhandensein von Itempositionseffekten wichtige Annahmen der IRT verletzen kann, ist es ein Ziel dieser Arbeit, eine Standardprozedur zur Berücksichtigung von Positionseffekten bei der Entwicklung eines computerisierten adaptiven Tests zu entwerfen. Im Projekt MaK-adapt wurde in einem ersten Schritt geprüft, ob Itempositionseffekte in den Daten vorliegen. Dazu wurde untersucht, wie häufig ein Item in Abhängigkeit von seiner Position im Testheft korrekt beantwortet wurde. Ein Streudiagramm bietet hierzu eine geeignete Möglichkeit die Positionseffekte darzustellen. Die Ergebnisse wurden domänenspezifisch betrachtet. Die Betrachtung der relativen Lösungshäufigkeiten bietet jedoch nur einen ersten Hinweis auf mögliche Itempositionseffekte. Zur Nutzung solcher Effekte müssen diese im Rahmen der IRT abgebildet werden. Da sich die bisherige Entwicklung der unidimensionalen Tests auf das Rasch-Modell bezieht, wurde sich zur Modellierung der Positionseffekte für ein Multifacetten-Rasch-Modell (vgl. Formel (12) auf S. 57) entschieden. So konnte das Multifacetten-Rasch-Modell zur Betrachtung der Positionseffekte mit dem einfachen Rasch-Modell (vgl. Formel (1) auf S. 18) ohne Berücksichtigung von Positionseffekten, nachfolgend auch Modell 1 genannt, verglichen werden. Ein möglicher Vergleichsaspekt war die Prüfung der Modelle auf deren Passung zu den Daten (globale Modellpassung z. B. über AIC oder BIC). Bei der Betrachtung von Positionseffekten über das Multifacetten-Rasch-Modell wurden in den nachfolgenden Analysen zwei Modelle unterschieden. Ein Modell betrachtet Itempositionseffekte, die für alle Items identisch sind (nachfolgend auch Modell 2 genannt) und das andere Modell betrachtet itemspezifische Positionseffekte (nachfolgend auch Modell 3 genannt; vgl. Formel (13) auf S. 57). D. h., bei Modell 3 können sich die Itempositionseffekte nicht nur zwischen den Positionen, sondern auch zwischen den unterschiedlichen Items unterscheiden. Dabei wurden die Modelle

bezüglich Modellpassung immer mit dem nächst komplexerem Modell verglichen (Modell 1 mit Modell 2 und Modell 2 mit Modell 3). Die Analyse der Positionseffekte erfolgte im Rahmen der IRT (Modell 2 und Modell 3) nicht auf den 33 Einzelpositionen eines Testheftes (vgl. Kapitel 4.3.2), sondern auf der Grundlage sogenannter Positionsstufen. Dabei wurden mehrere Einzelpositionen zu einer Positionsstufe zusammengefasst. Auf diesem Weg konnten mehr Datenpunkte (Probanden pro Item pro Position) für eine stabilere Schätzung der Positionseffekte genutzt werden. Bei Modell 3 wäre ohne diese Aggregation eine Schätzung der itemspezifischen Positionseffekte nicht möglich gewesen, da die Anzahl der Probanden auf einem Item an einer Position teilweise sehr gering war. In der Domäne Lesen wurden drei Positionen und in den Domänen Mathematik und Naturwissenschaft jeweils fünf Positionen zu einer Positionsstufe zusammengefasst. So ergaben sich für Lesen neun und für Mathematik und Naturwissenschaft jeweils sieben Positionsstufen. In der Domäne Lesen wurde die Auflösung (Anzahl an Positionen in einer Positionsstufe) kleiner gefasst, da dort aufgrund des Testheftdesigns jedes Testheft häufiger vorgelegt werden konnte und somit die Anzahl an Antworten pro Position größer war, als bei Mathematik und Naturwissenschaft (Frey et al., im Druck).

Im Durschnitt ergibt sich durch die Bildung der Positionsstufen eine mittlere Anzahl an Antworten pro Position für die Domäne Lesen von $N_{\text{mean}} = 31.176$ ($N_{\text{min}} = 12$), für die Domäne Mathematik von $N_{\text{mean}} = 34.366$ ($N_{\text{min}} = 11$) und für die Domäne Naturwissenschaft von $N_{\text{mean}} = 34.213$ ($N_{\text{min}} = 11$). Bei Betrachtung jeder einzelner Position wären durchschnittlich für die Domäne Lesen $N_{\text{mean}} = 10.392$ ($N_{\text{min}} = 3$), für die Domäne Mathematik $N_{\text{mean}} = 7.300$ ($N_{\text{min}} = 3$) und für die Domäne Naturwissenschaft $N_{\text{mean}} = 7.761$ ($N_{\text{min}} = 1$) Antworten pro Position erreicht worden. Für jede Domäne wurden insgesamt drei unterschiedliche unidimensionale Modelle aufgrund deren Devianzen und durch Informationskriterien miteinander über einen Likelihood-Quotienten-Test verglichen. Die geschätzten Itemschwierigkeiten aus den verschiedenen Modellen wurden über einen Chi-Quadrat-Differenzentest verglichen. Dazu wurde (a) die Differenz der Devianzen von Modell 2 zu Modell 1 und (b) von Modell 3 zu Modell 2 verglichen. Anschließend wurden die Informationskriterien BIC, AIC und CAIC miteinander verglichen. Für die Signifikanztests wurde immer ein zweiseitiger Test mit einem Signifikanzniveau von $\alpha = .05$ durchgeführt. Die Frage nach der Auswirkung der Modellierung von Positionseffekten auf die Varianz und die Reliabilität der Personenver-

teilungen wurde beantwortet, indem die latenten Varianzen und die EAP/PV-Reliabilitäten (Adams, 2005) der Personenverteilung für die oben beschriebenen Modelle geschätzt und miteinander verglichen wurde. Für eine Vergleichbarkeit wurde bei der Schätzung der unterschiedlichen Modelle darauf geachtet, dass der Mittelwert der Personenverteilung bei der Schätzung auf 0 fixiert wurde. Die Schätzung der verschiedenen Modelle erfolgte mithilfe der Software ConQuest. Die Spezifikation der Modelle erfolgte mit den nachfolgenden Befehlen:

Modell 1: `model Item;`

Modell 2: `model Item + Position;`

Modell 3: `model Item + Position + Item*Position;`

Bei acht der neun geschätzten Modelle zeigten sich keine Konvergenzprobleme. Für die Domäne Mathematik konnte beim Modell 3 für einige Items die itemspezifischen Positionseffekte nicht identifiziert werden. Diese Parameter wurden von der Schätzung ausgeschlossen. Zusätzlich zu den beschriebenen unidimensionalen Skalierungen wurden die Ergebnisse aus der Kalibrierungsstudie für die drei Domänen als multidimensionales Modell skaliert. Auf diese Weise konnten alle 33 Items in einem Testheftzusammen in die Schätzung eingehen. So ergab sich ein Hinweis über mögliche Positionseffekte innerhalb der Bearbeitung eines gesamten Testheftes über alle Positionsstufen und drei Domänen hinweg. Bei der multidimensionalen Schätzung wurden drei Positionen zu einer Positionsstufe zusammengefasst. So ergeben sich 11 Positionsstufen wurden. Es wurde nur ein Modell geschätzt, bei dem die Positionseffekte für alle Items als identisch angesehen werden (vgl. Modell 2). Jedes Item wurde genau einer Domäne zugeordnet. In ConQuest wird die Syntax zur Festlegung der Dimensionen mit der ersten Dimension für 94 Naturwissenschaftsitems, der zweiten Dimension für 105 Mathematikitems und der dritten Dimension für 65 Leseitems folgendermaßen geschrieben:

```
score (0 1) (0 1) () () !items(1-94);

score (0 1) () (0 1) () !items(95-199);

score (0 1) () () (0 1) !items(200-264);
```

Einen ersten Einblick, ob Itempositionseffekte in den Daten vorliegen, liefert die nachfolgende Abbildung 11.

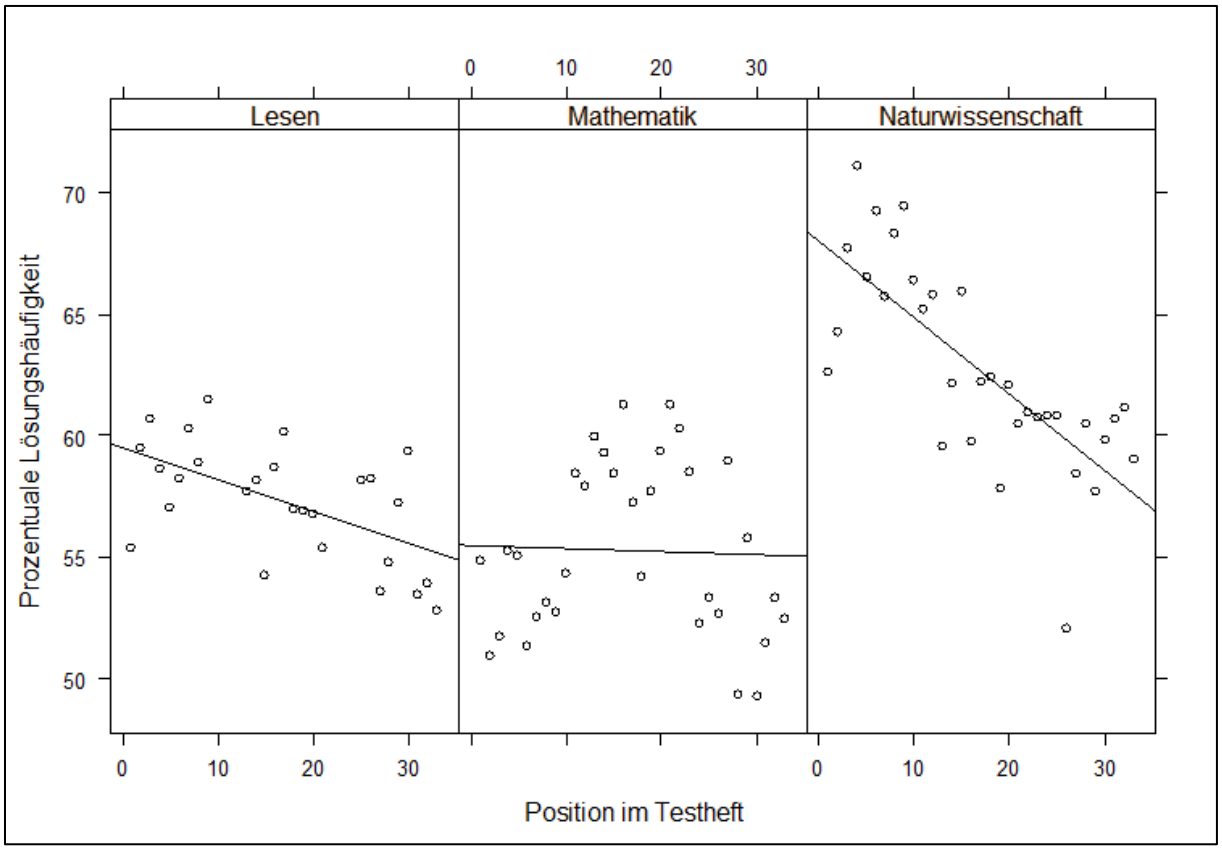

Abbildung 11. Mittlere relative Lösungshäufigkeit aller auf der gleichen Position vorgegebenen Items nach Position und Domäne.

In der Abbildung ist die prozentuale Lösungshäufigkeit aller Items in Abhängigkeit von der Darbietungsposition im Testheft je Domäne abgebildet. Für die Domäne Mathematik ist sichtbar, dass ein lineares Modell die Daten nicht gut abbilden würde. Es kann anhand der Abbildung interpretiert werden, dass die Positionseffekte domänenspezifisch ausfallen. Die Lösungshäufigkeiten fallen in den Domänen Lesen und Naturwissenschaft im Verlauf des Testhefts ab. In der Domäne Mathematik sind auf den mittleren Positionen die höchsten Lösungshäufigkeiten und am Anfang bzw. am Ende des Testheftes niedrigere Lösungshäufigkeiten zu beobachten. In den Domänen Lesen und Naturwissenschaft können die Itempositionseffekte durch lineare Trends gut erklärt werden. Die Lösungshäufigkeiten in der Domäne Mathematik lassen sich besser mit einem quadratischen Trend erklären. In der Tabelle 8 sind die Ergebnisse zur globalen Modellpassung für die drei Modelle abgebildet.

Tabelle 8

Globale Modellpassung für Modell 1, Modell 2 und Modell 3 für die Tests im Lesen (READ), in der Mathematik (MATH) und in der Naturwissenschaft (SCIE)

Domäne	Modell	Deviance	*m*	*p*	AIC	CAIC	BIC
READ	1	14 213.12	66		14 345.12	14 350.77	14 701.36
	2	14 123.49	74	< .001	14 271.49	14 278.62	14 670.91
	3	13 621.09	584	.586	14 789.09	15 441.70	17 941.27
MATH	1	15 483.43	106		15 695.43	15 710.31	16 267.57
	2	15 376.34	112	< .001	15 600.34	15 617.00	16 204.86
	3	14 795.58	723	.805	16 241.58	17 394.56	20 144.02
SCIE	1	14 006.52	95		14 196.52	14 208.39	14 709.29
	2	13 874.83	101	< .001	14 076.83	14 090.30	14 621.98
	3	13 333.87	650	.588	14 633.87	15 496.56	18 142.28

Anmerkungen. Datengrundlage: $N = 1\,632$ Probanden. Deviance = 2*log-Likelihood. Modell 1: Rasch-Modell. Modell 2: Multi-Facetten-Rasch-Modell mit itemunspezifischen Positionseffekten. Modell 3: Multi-Facetten-Rasch-Modell mit itemunspezifischen und itemspezifischen Positionseffekten. m: Anzahl Modellparameter, p: Irrtumswahrscheinlichkeit Chi-Quadrat-Differenzentest zum Modellvergleich mit weniger komplexem Modell in vorheriger Zeile.

Das Modell 2 zeigt für alle drei Domänen eine signifikant bessere Modellpassung ($p < .001$) als das Modell 1. Das Modell 3 passt in keinen der Domänen besser als Modell 2. Die Informationskriterien AIC, CAIC und BIC sprechen ebenfalls in allen drei Domänen für das Modell 2. Die Schätzung von Populationskennwerten erfolgte über eine latente Populationsverteilung und wurde mittels numerischer Verfahren approximiert. Dazu wurde eine diskrete Anzahl von Knoten (Nodes) über die latente Merkmalsskala verteilt und die Dichtefunktion über jedem Knoten aus den empirischen Daten berechnet. Die nachfolgenden Ergebnisse unterscheiden sich minimal von den Ergebnissen bei Frey et al. (im Druck), da bei den Schätzungen eine unterschiedliche Anzahl von Nodes ver-

wendet wurde. Diese Änderungen in der dritten Nachkommastelle haben jedoch keinen Einfluss auf die inhaltlichen Ergebnisse. Insgesamt lassen sich in allen drei Domänen Itempositionseffekte finden. Die Modellpassung von Modell 3 für die Domäne Lesen ($p = .586$), für die Domäne Mathematik ($p = .805$) und für die Domäne Naturwissenschaft ($p = .588$) im Vergleich zu Modell 2 fallen relativ schlecht aus. Die Informationskriterien AIC, CAIC und BIC sprechen ebenfalls bei allen drei Domänen gegen das Modell 3. Das Modell 2 ist im Vergleich zu Modell 3 zudem wesentlich sparsamer und wird hier deshalb für alle drei Domänen als endgültiges Modell gewählt. Es ist entsprechend festzuhalten, dass die Positionseffekte für alle Items einer Domäne identisch sind und keine itemspezifischen Positionseffekte vorliegen.

Um die Frage zu beantworten, wie sich die Modellierung von Itempositionseffekten auf die Varianz und die Reliabilität der gemessenen Merkmalsausprägungen auswirkt, wurden die Ergebnisse dazu für das gewählte Modell 2 mit den Ergebnissen aus Modell 1 verglichen. Die Hinzunahme itemunspezifischer Positionseffekte bei Modell 2 führt dabei zu kleinen Verringerungen von Varianz und Reliabilität im Vergleich zu Modell 1. Für Lesen verringert sich die Varianz von 0.739 auf 0.735 und die Reliabilität von 0.487 auf 0.484. Für Mathematik verringert sich die Varianz von 0.954 auf 0.935 und die Reliabilität von 0.552 auf 0.546 und für Naturwissenschaft verringert sich die Varianz von 0.763 auf 0.721 und die Reliabilität von 0.478 auf 0.463. Im Ergebnis lässt sich feststellen, dass die Modellierung der Itempositionseffekte zu keinen nennenswerten Einschränkungen hinsichtlich der Varianz und Reliabilität der individuellen Merkmalsschätzer führt. Anders ausgedrückt kann auch festgestellt werden, dass die vorhandenen und mitmodellierten Positionseffekte nicht anhand der Personenverteilung sichtbar werden. In Bezug auf die Größe der Itempositionseffekte wurde zuerst die multidimensionale Skalierung für Modell 2 über 11 Positionsstufen und die drei Domänen hinweg geprüft. Durch die multidimensionale Skalierung lagen im Vergleich zur unidimensionalen Skalierung mehr Daten für die Schätzung der Positionseffekte vor, wodurch kleinere Standardfehler erreicht werden konnten ($\overline{SE} = 0.030$). Zudem konnte dadurch ein Bild der Positionseffekte über das gesamtes Testheft und alle drei Domänen hinweg gezeichnet werden. Denn auch wenn von domänenspezifischen Positionseffekten ausgegangen werden kann, bekam bei der Kalibrierungsstudie doch jeder Proband alle drei Domänen in einem Testheft vorgelegt. In der Abbildung 12 sind die Effekte für die multidimensio-

nale Skalierung abgebildet. Es ist ein Anstieg der Effekte von der zweiten Positionsstufe mit -0.096 Logits bis zur letzten Stufe mit 0.176 Logits zu sehen. Das entspricht einem Abstand von 0.272 Logits über das Testheft hinweg. Die Positionsstufen 4 und 8 weisen Sprünge nach unten auf. An diesen Stufen wurden aufgrund des Testheftdesigns keine Leseitems vorgegeben. Deshalb sollten die Ergebnisse an diesen Stufen gesondert interpretiert werden.

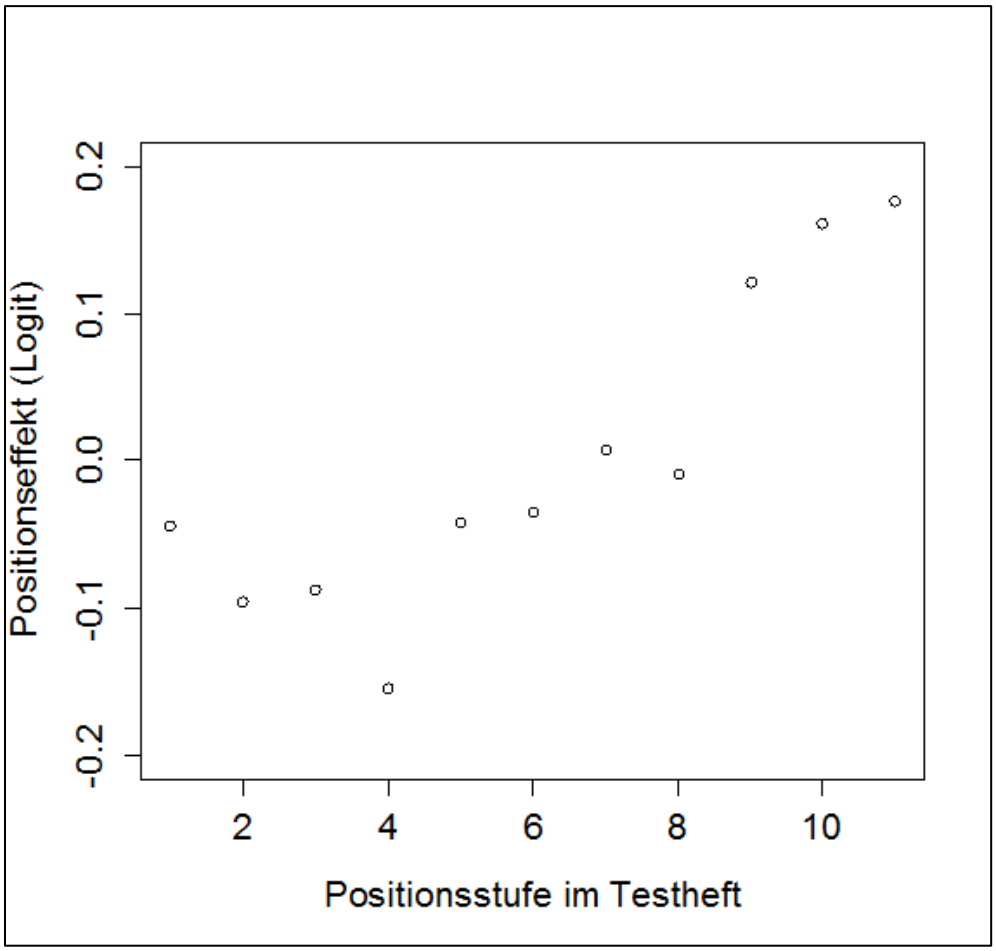

Abbildung 12. Positionseffekte in Logits aus Modell 2 für die multidimensionale Skalierung.

In der nachfolgenden Tabelle 9 sind die Positionseffekte bei eindimensionaler Skalierung für jede Domäne zu sehen. Die Positionseffekte mit den maximalen Werten 0.220 bei Lesen, -0.152 bei Mathematik und -0.267 bei Naturwissenschaft erscheinen auf den ersten Blick vertretbar. Dennoch muss berücksichtigt werden, dass diese Effekte im Testverlauf auf mehrere Items zutreffen und sich somit potenzieren, was die Auswahl der Items sowie die Personenparameterschätzung beeinflussen kann.

Tabelle 9

Positionseffekte und Standardfehler (SE) in den Domänen Lesen (READ), Mathematik (MATH) und Naturwissenschaft (SCIE)

Domäne	Positionsstufe	Positionseffekt	SE
READ	1	-0.077	0.045
	2	-0.074	0.045
	3	-0.160	0.046
	4	-	-
	5	0.022	0.049
	6	-0.070	0.050
	7	0.074	0.050
	8	-	-
	9	0.037	0.052
	10	0.027	0.053
	11	0.220*	0.138
MATH	1	0.051	0.046
	2	0.111	0.044
	3	-0.152	0.041
	4	-0.078	0.044
	5	-0.100	0.042
	6	0.093	0.046
	7	0.074*	0.107

Domäne	Positionsstufe	Positionseffekt	SE
SCIE	1	-0.174	0.048
	2	-0.267	0.046
	3	-0.118	0.043
	4	0.072	0.046
	5	0.076	0.042
	6	0.252	0.045
	7	0.160*	0.110

Anmerkungen. Aufgrund des Testheftdesigns konnte an den Positionsstufen vier und acht kein Leseitem vorgelegt werden. Der letzte mit einem Stern markierte Positionseffekt in jeder Domäne wurde nicht frei geschätzt, sondern so fixiert, dass die Summe aller Werte 0 ergibt.

Weiterhin hat sich gezeigt, dass sich bei der Berücksichtigung von Itempositionseffekten mit Modell 2 die Itemschwierigkeiten im Vergleich zu Modell 1 ändern können. In der Domäne Mathematik ändert sich bei 16 Items die Schwierigkeit um mehr als 0.1 Logits. Dabei handelt es sich ausschließlich um Items, welche in Randbereichen der Verteilung der Itemschwierigkeiten liegen. Da dort die Standardfehler der Parameterschätzer besonders hoch sind, ist eine erhöhte Differenz nicht ungewöhnlich. Dennoch liegt bei dem leichtesten Mathematikitem (Item 57 in der nachfolgenden Wright-Map; vgl. Abbildung 13) ein Unterschied von 0.829 Logits vor. Das Item hat somit bei der Skalierung der Daten ohne Positionseffekt eine Schwierigkeit von -4.204 ($SE = 0.591$) und bei der Skalierung mit Positionseffekt nur noch eine Schwierigkeit von -3.375 ($SE = 0.606$). Leichte Items werden nach der Berücksichtigung der Positionseffekte schwieriger und schwierige Items werden leichter. Hierzu wurde nachfolgend beispielhaft die Wright-Map für die Domäne Mathematik abgebildet, wobei hier nur die Auflistung der Items (a) ohne Berücksichtigung von Positionseffekten und (b) mit Berücksichtigung von Positionseffekten erfolgt. Die Verteilung der Personen wurde aus Gründen der Übersichtlichkeit nicht mit abgebildet.

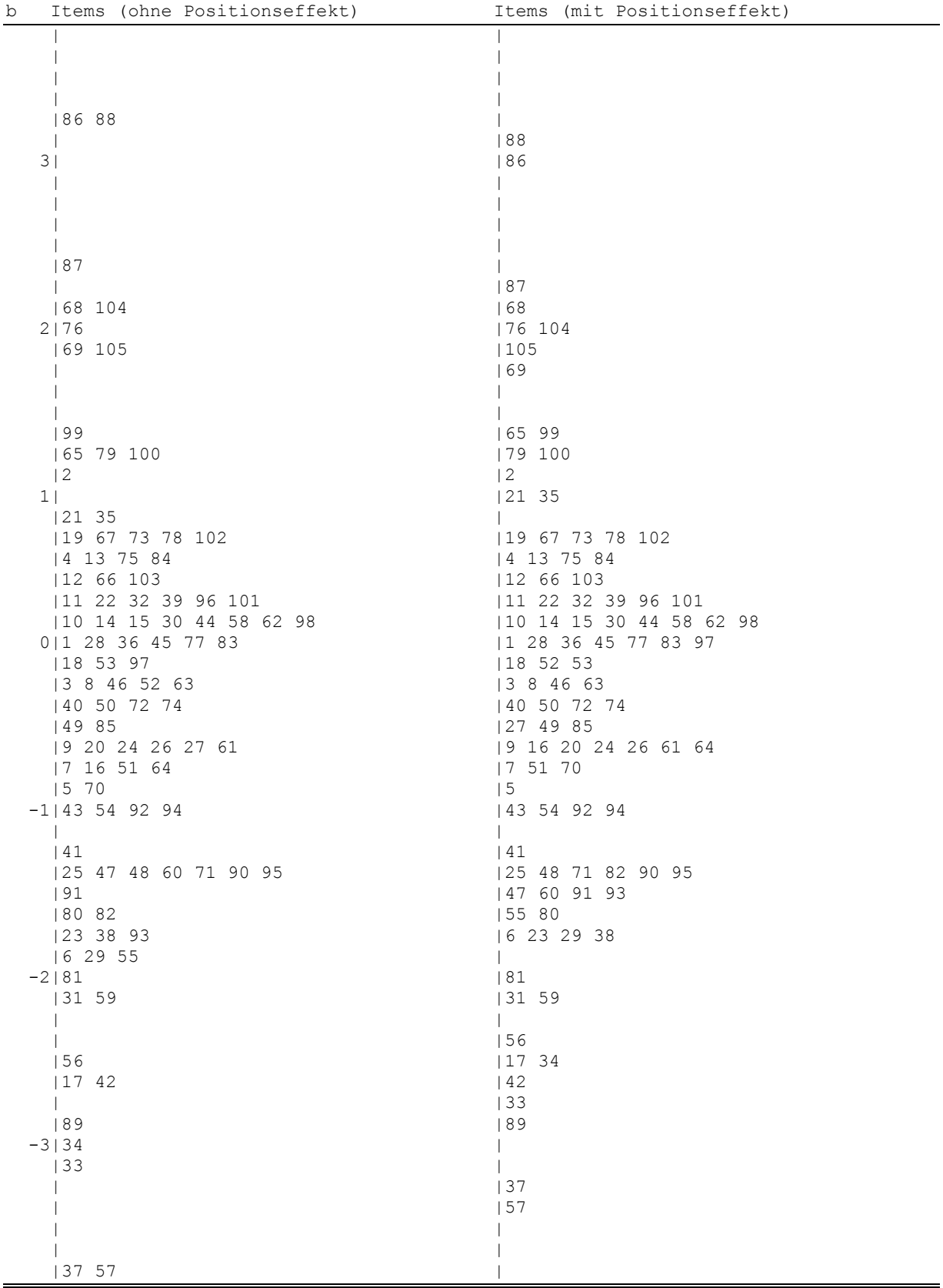

Abbildung 13. Wright-Map zur Verteilung der Items über den Schwierigkeitsbereich für die Skalierung ohne Positionseffekte (Modell 1) und mit Positionseffekten (Modell 2) für die Domäne Mathematik.

In der Domäne Lesen ändern sich acht Items um mehr als 0.1 Logits (von 0.120 bis 0.479). Alle acht Items stammen aus dem unteren Bereich der Schwierigkeitsverteilung (im Schwierigkeitsbereich von -0.981 bis -3.001 bei der Skalierung ohne Positionseffekte). Alle acht leichten Items werden bei Betrachtung von Positionseffekten schwieriger. Nur ein Item wird leichter (um -0.062 Logits). Dieses stammt aus dem Bereich der eher schwereren Items. In der Domäne Naturwissenschaft ändern sich 18 Items um mehr als 0.1 Logits (von 0.109 bis 0.472). Alle 18 Items stammen wie bei Lesen aus dem unteren Bereich der Schwierigkeitsverteilung (im Schwierigkeitsbereich von -0.666 bis -2.984 bei Skalierung ohne Positionseffekte). Alle 18 leichten Items werden bei Betrachtung von Positionseffekten schwerer.

4.3.6 Zusammenfassung

Es wurde in der Kalibrierungsstudie ein komplexes balanciertes Testheftdesign mit 798 unterschiedlichen Testheften zur Kalibrierung verwendet, mit dem Ziel, alle Items gleichmäßig über die Positionen im Testheft hinweg zu verteilen. Dieses Vorgehen ermöglicht später die Schätzung von Schwierigkeitsparametern auf Positionsebene. Dadurch, dass ein unvollständiges Design gewählt wurde, konnten trotz der hohen Anzahl an zu kalibrierenden Items kurze Testhefte mit 33 Items konstruiert werden. Die Testhefte in der Kalibrierungsstudie wurden spiralisiert in den Klassen verteilt. Nach der Datenaufbereitung gingen die Ergebnisse von 1,632 SuS in die Kalibrierungsergebnisse ein. Dabei wurde darauf geachtet, die SuS so zu wählen, dass die ASCOT-Berufe ausreichend vertreten sind, damit die Stichprobe der Kalibrierung ähnlich der Personen ist, an denen die Tests später verwendet werden. Aufgrund der vorhandenen Daten und der Kalibrierung der Items für einen unidimensionalen adaptiven Test wurde sich dafür entschieden, die Daten für jede Domäne mit einem eindimensionalen Modell für dichotome Daten, dem Rasch-Modell, zu skalieren. Anschließend wurden aufgrund der Skalierungsergebnisse und der Betrachtung statistischer und inhaltlicher Kriterien defizitäre Items identifiziert und aus dem Itempool entfernt. Dabei wurde besonders auf die Trennschärfe, den Itemfit und DIF geachtet. Es mussten dabei nur wenige Items aufgrund von DIF entfernt werden. Die Anzahl der Items ist innerhalb der Domänen über die Inhaltsbereiche annähernd gleichverteilt. Die Verteilung der Schwierigkeitsparameter ist für einen adaptiven Test, der auch in den Randbereichen der Kompetenzverteilung

optimal funktionieren soll, jedoch verbesserungswürdig. Hier können Items, in den Rändern des Schwierigkeitsbereichs nachträglich dem Itempool hinzugefügt werden.

Die Überprüfung auf Itempositionseffekte ergab, dass Positionseffekte vorliegen, die für alle Items als identisch angesehen werden können (itemunspezifisch). In der Studie von Frey et al. (im Druck) kam zusätzlich das Ergebnis heraus, dass für die Domäne Naturwissenschaft ein etwas komplexeres Modell besser passt, bei dem itemspezifische Positionseffekte vorliegen. Dabei bezieht sich die Itemspezifität auf die Länge der Items (Anzahl der Wörter der Items). Die hier vorgestellten itemunspezifischen Positionseffekte sind dennoch auch für die Domäne Naturwissenschaft eine bessere Annährung als das Ignorieren dieser Effekte. Die Effekte können als Resultat einfach und direkt beim CAT genutzt werden. Das beschriebene Vorgehen kann als Routineverfahren umgesetzt werden. Die relative Lösungshäufigkeit in den Domänen Lesen und Naturwissenschaft fielen gegen Ende des Tests ab. Im Bereich Mathematik zeigt sich zu Beginn und am Ende der Testung eine geringe relative Lösungshäufigkeit. In der Mitte der Testung hingegen steigt diese Lösungshäufigkeit an. Als passendes Modell zur Modellierung wurde das Modell 2 mit identischen Positionseffekten für alle Items einer Domäne gewählt. Das bedeutet, dass jedes Item seine Schwierigkeit über den Test hinweg in gleicher Weise ändert. Eine mögliche Erklärung wäre, dass die Items so konstruiert wurden, dass sie bezüglich Darbietung relativ ähnlich zueinander sind. Die Verwendung von heterogenerem Itemmaterial, wie es in anderen Studien teilweise angewendet wird, könnte dazu führen, dass die Modellierung von itemspezifischen Positionseffekten notwendig wird. Dennoch können auch dann Itemparameter und Positionsparameter direkt für CAT genutzt werden. Weiterhin haben die Ergebnisse gezeigt, dass die Modellierung der Positionseffekte mit der Modellierung eines für alle Items identischen Positionseffekts, nur marginale Auswirkungen auf die Varianz und Reliabilität der Skalen hat. Hierbei ist anzumerken, dass die Reliabilität insgesamt gering ausfällt, da der Fokus der Kalibrierung auf der Schätzung der Itemparameter lag und deshalb jedem SuS nur wenige Items vorgelegt wurden (Lesen 7 Items; Mathematik und Naturwissenschaft jeweils 12 Items). Die vorgestellten Qualitätskontrollen in Bezug auf die Entwicklung eines CAT-Itempools versprechen, dass aus bestehenden Items faire Testinstrumente mit hoher Passung zum Rasch-Modell konstruiert werden können. Weiterhin wurde verdeutlicht, dass vorliegende Positionseffekte sehr komplex sein können und inhaltlich viele Begründungen

zulassen. Deshalb ist es wichtig, ein einfaches Modell zur Modellierung zu verwenden, was zugleich sehr flexibel ist. Das Multifacetten-Rasch-Modell wird deshalb an dieser Stelle empfohlen.

4.4 CAT – Algorithmus

In diesem Abschnitt werden die einzelnen Schritte zur Festlegung des adaptiven Algorithmus entsprechend des Pfaddiagramms zum Ablauf adaptiver Tests (vgl. Abbildung 2 auf S. 59) festgelegt. D. h., der Startpunkt, der Itemauswahlmechanismus, die Methode der Fähigkeitsschätzung und das Abbruchkriterium für die empirisch im Projekt MaK-adapt entwickelten Tests werden in diesem Abschnitt spezifiziert und die getroffenen Entscheidungen erläutert. Zudem wird auf zusätzliche Restriktionen bei der Itemauswahl, dem Ausbalancieren der Inhaltsbereiche aus dem inhaltlichen Zielkonstrukt (Content-Balancing) eingegangen. Fragen, die auf die Funktionsweise des Algorithmus im Zusammenspiel mit dem Itempool abzielen, werden u. a. durch Simulationsstudien beantwortet.

4.4.1 Fragestellungen

- Wie wird der Personenparameterschätzer zum Teststart spezifiziert?
- Wie erfolgt die Personenparameterschätzung während der Testung?
- Wie erfolgt die Itemauswahl zu Beginn der Testung?
- Wie erfolgt die Itemauswahl während der Testung?
- Nach welchen Kriterien wird der Test beendet?
- Erfüllt die Content-Balancing-Methode die Balancierung der Subdimensionen in diesem Itempool angemessen?
- Wie hoch ist die zu erzielende Messpräzision in den einzelnen Domänen bei der Nutzung des Itempools und der Verwendung des MPI (Simulationsstudien)?

4.4.2 Methode und Ergebnisse: Algorithmus festlegen

Bei der Festlegung des Algorithmus wurde davon ausgegangen, dass zu Beginn der Testung keine Kenntnisse über die Fähigkeit der Testperson vorliegen. Die beste

Annahme unter dieser Voraussetzung besteht darin, für alle getesteten Personen anzunehmen, dass ihre Fähigkeit für jeden der drei gemessenen Kompetenzbereiche (Mathematik, Naturwissenschaft und Lesen) dem Mittelwert der Kalibrierungsstichprobe entspricht. Dieser wurde in allen drei Domänen auf den Wert 0 fixiert. D. h., die Fähigkeitsschätzer jeder Person besitzen zu Beginn der Testung den Wert 0. Die Itemauswahl erfolgt nach maximaler Information (vgl. Formel (14) auf S. 61). Dieses Vorgehen hat den Nachteil, dass alle Personen zu Beginn der Messung das gleiche Item vorgegeben bekommen, sofern es für den Fähigkeitsschätzer 0 nur ein Item mit maximaler Information gibt. Um dies zu verhindern, wurde per Zufall aus 10 Items mit hoher Information ein Startitem ausgewählt. Zur Fähigkeitsschätzung wird der BME (vgl. Formel (19) auf S. 62) verwendet, da dieser auch bei kurzen Tests, in dem alle Items korrekt bzw. falsch beantwortet wurden, eine Schätzung liefert. Da der BME die a-priori-Verteilung der Fähigkeit der Kohorte aus den Daten der Kalibrierungsstudie berücksichtigt, ist bei kurzen Testungen zudem ein präziseres Ergebnis als bei beispielsweise dem MLE zu erwarten. Als Abbruchkriterien wurden die Parameter Testlänge (maximale Itemanzahl) und Testzeit verwendet. Die Testlänge beschreibt die Anzahl an vorgelegten Items innerhalb eines Tests und wurde verwendet, um sicherzustellen, dass jeder Proband dieselbe Anzahl an Items erhält. Wenn die Anzahl an Items erreicht ist, wird der Test beendet. Um die Testzeit für alle Probanden gleich zu halten, kann dieses als Kriterium verwendet werden. Dies wurde hier ebenfalls verwendet, da in den Schulen maximale Zeiten zur Testung zur Verfügung gestellt wurden. Zur Zeitmessung dient die interne Uhr des Computers. Die Software MATE beginnt dabei mit der Zeitmessung ab Beginn eines Tests. Wenn die vorher definierte Zeitgrenze erreicht wurde, stoppt der Test und die Testung ist beendet. Dabei wird von der Software MATE folgende Meldung ausgegeben, um mit dem Probanden in Interaktion zu treten: *The time for this part of the test is over. Please click on OK to proceed.* Die eingestellten Standardwerte zur Testbeendigung für die MaK-adapt Pilotierungsstudie sind eine Testzeit von maximal 40 Minuten und eine maximale Testlänge von 48 Items. Nach 48 Items wird eine hinreichend große Reliabilität erwartet (vgl. Abbildung 16 auf S. 136).

Das Content-Balancing erfolgte im adaptiven Algorithmus über die Methode des MPI (vgl. Formel (24)auf Seite 66). Dabei kann unterschieden werden, ob der MPI die absoluten Anteile (MPI 1) oder die relativen Anteile an vorgegebenen Items (MPI 2)

kontrolliert. Im eingestellten Algorithmus für die Pilotierungsstudie wurde nach absoluten Anteilen kontrolliert. Konkret wurde die festgelegte Anzahl an maximalen Items pro Domäne gleichmäßig auf die Subdomänen (Inhaltsbereiche) verteilt. Die Werte für die maximale Anzahl an vorzulegenden Items je Inhaltsbereich sind der Tabelle 10 zu entnehmen.

Tabelle 10

Content-Balancing-Restriktionen pro Inhaltsbereich für die Tests im Lesen (READ), in der Mathematik (MATH) und in der Naturwissenschaft (SCIE)

Domäne	Inhaltsbereich	Anzahl Items
READ	Deskriptionale Darbietung	16
	Gemischte Darbietung	16
	Depiktionale Darbietung	16
MATH	Quantität	12
	Veränderung und Beziehung	12
	Raum und Form	12
	Unsicherheit	12
SCIE	Leben und Gesundheit	12
	Erde, Planeten, Umwelt und natürliche Ressourcen	12
	Stoffe und Stoffveränderungen	12
	Bewegung, Kraft und Energie	12

Die aufgeführten Einstellungen für den computerisierten adaptiven Algorithmus wurden auf Grundlage von Simulationsstudien festgelegt. Das Ziel der Simulationsstudien war es, (a) die zu erwartende Messpräzision (Reliabilität; vgl. Kapitel 3.5.4) in Zusammenhang mit dem Abbruchkriterium Itemanzahl festzustellen und (b) die gleichmäßige Vorgabe von Items aller Subdomänen durch den MPI zu prüfen. Zudem wurde in der Simulationsstudie die Messpräzision des computerisierten adaptiven Tests mit Benutzung des MPI im Vergleich zum FIT geprüft. Dabei wurde die Schätzung des zu erwartenden Standardfehler (*SE*; vgl. Formel (20) auf S. 64) und der zu erwartenden

marginalen Reliabilität zwischen FIT und CAT in Abhängigkeit von θ verglichen. Die Simulationen erfolgte in einem ersten Schritt in der Software SAS 9.3. Es wurden 50 Replikationen und $N = 1\,000$ Probanden gewählt. Die angenommene Verteilung der Probanden war eine Normalverteilung mit einem Mittelwert von 0 und einer Varianz von 1. Für die Prüfung des Standardfehler (*SE*) in Abhängigkeit von θ wurden die Ergebnisse nach 32 Items (Domäne Lesen) bzw. 36 Items (Domänen Mathematik und Naturwissenschaft) gewählt. Diese Testlängen wurden gewählt, da später beim FIT eine ähnliche Testlänge angestrebt wurde und somit die Ergebnisse gut vergleichbar sind (vgl. endgültige Wahl der FIT-Testlänge in Kapitel 4.6). In der nachfolgenden Abbildung ist das Ergebnis für den *SE* in Abhängigkeit von θ für die Domäne Mathematik zu sehen. Es wird deutlich, dass Standardfehler, vor allem im Randbereich von θ, beim FIT deutlich höher sind als beim CAT.

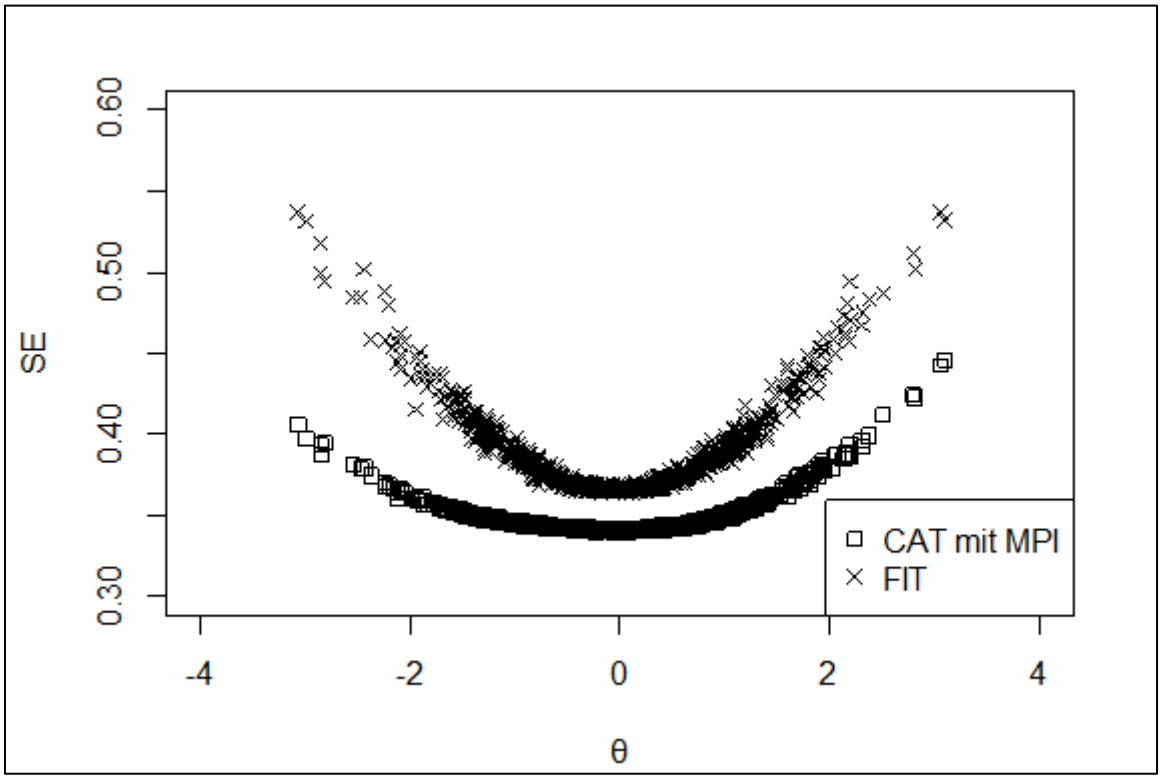

Abbildung 14. Standardfehler (*SE*) vom Personenparameterschätzer nach 36 Items in Abhängigkeit von der Merkmalsausprägung (θ) im Vergleich von FIT und CAT mit MPI für die Domäne Mathematik.

Für die Domänen Naturwissenschaft und Lesen (vgl. Abbildung 15 und Abbildung 16) fällt der *SE* vor allem bei den Personen mit hohen Werten für θ höher aus. Dies ist darauf

zurückzuführen, dass in diesem Schwierigkeitsbereich relativ wenig Items im Itempool vorhanden sind. Der Itempool von Naturwissenschaft enthält hingegen sehr viele leichte Items. Dadurch wird im linken Randbereich (negativer Bereich der Verteilung von θ) ein vergleichbar geringer *SE* erzielt, wie in der Mitte der Verteilung.

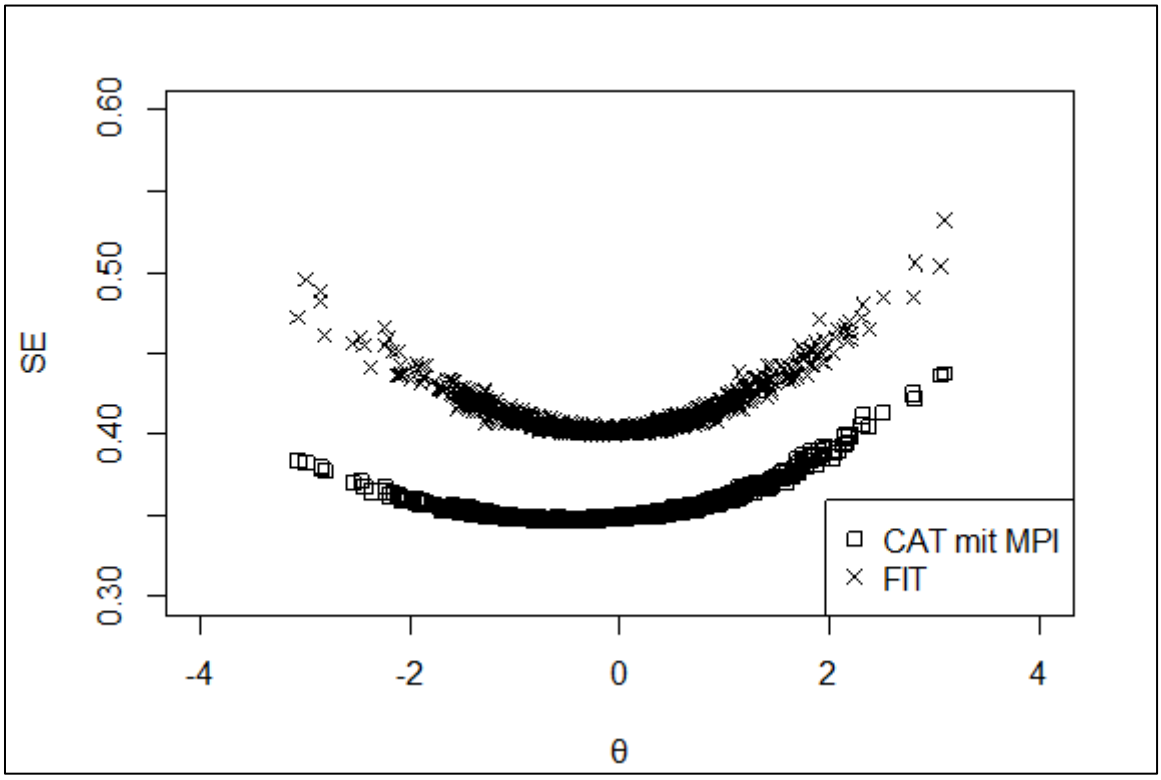

Abbildung 15. Standardfehler (*SE*) vom Personenparameterschätzer nach 32 Items in Abhängigkeit von der Merkmalsausprägung (θ) im Vergleich von FIT und CAT mit MPI für die Domäne Lesen.

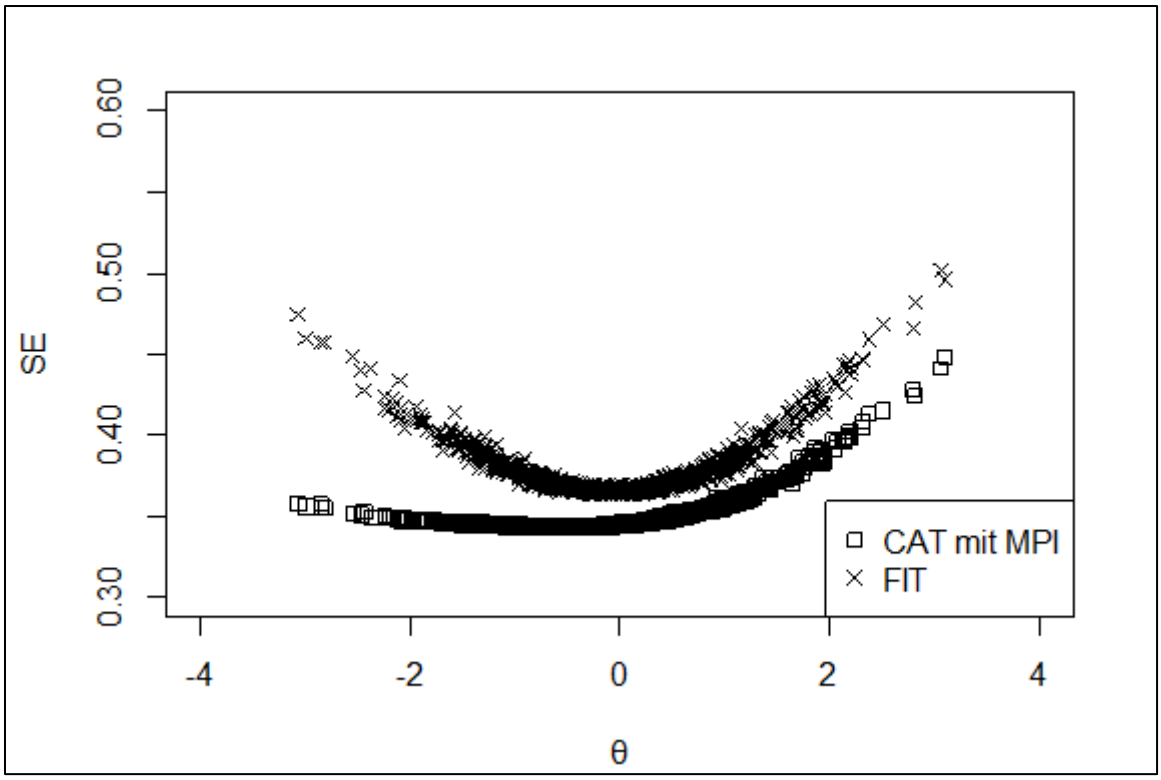

Abbildung 16. Standardfehler (*SE*) vom Personenparameterschätzer nach 36 Items in Abhängigkeit von der Merkmalsausprägung (θ) im Vergleich von FIT und CAT mit MPI für die Domäne Naturwissenschaft.

Die zu erwartende Messpräzision (Reliabilität) für CAT in Abhängigkeit von der Testlänge (Anzahl Items) der drei Domänen ist in der Abbildung 17 zu sehen. Die Reliabilität wird in der Abbildung erst ab einer Testlänge von neun Items (Domäne Mathematik) bzw. 10 Items (Domänen Lesen und Naturwissenschaft) angegeben, da sich bei der verwendeten Reliabilitätsberechnung (vgl. Formel (22) auf S. 65) bei geringerer Testlänge negative Werte ergeben. Die maximale Anzahl von 48 Items in der Abbildung wurde gewählt, da dies einem Abbruchkriterium im Algorithmus für die Pilotierungsstudie entspricht. In der Abbildung wird deutlich, dass die Domänen Lesen und Naturwissenschaft einen vergleichbaren Verlauf der zu erwartenden Reliabilität aufweisen; der Mathematiktest weist bei gleicher Testlänge eine höhere Reliabilität auf.

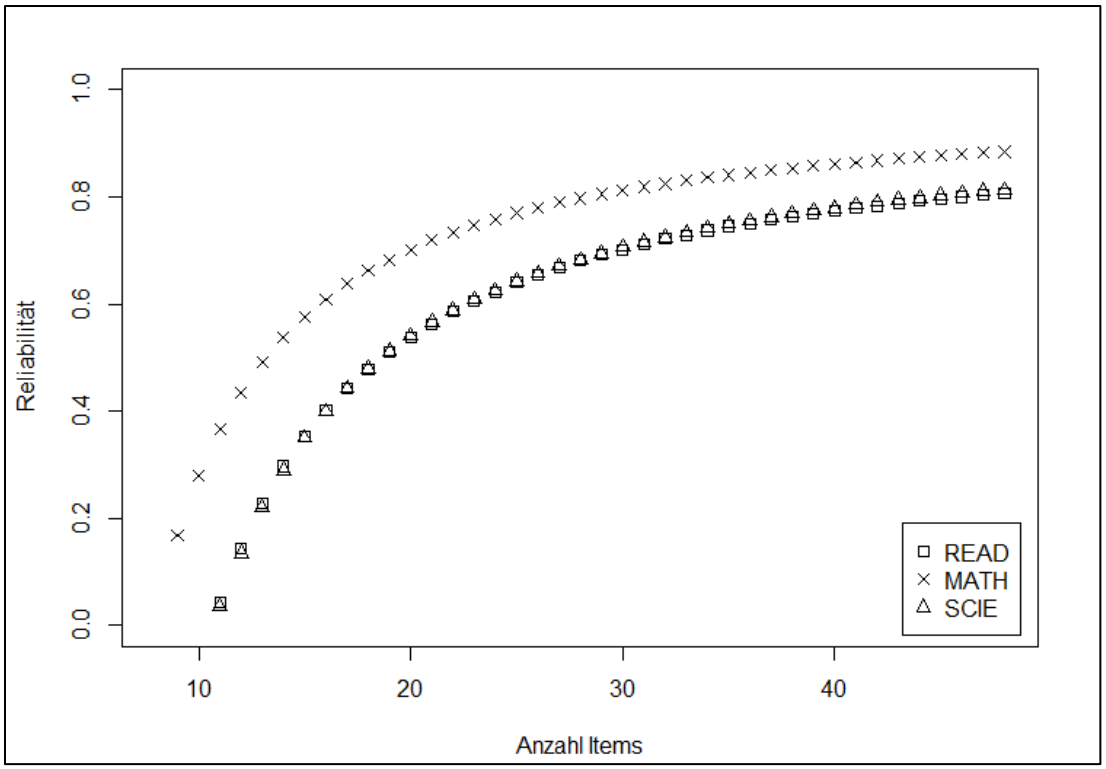

Abbildung 17. Marginale Reliabilität für CAT in Abhängigkeit von der Anzahl vorgegebener Items für eine maximale Testlänge von 48 Items für die Domänen Lesen (READ), Mathematik (MATH) und Naturwissenschaft (SCIE).

Eine Reliabilität von über .7 wird beim computerisierten adaptiven Testen in der Domäne Mathematik bereits ab 20 Items ($\rho(\hat{\theta}_u\theta_u)^2 = .701$) erreicht. In der Domäne Lesen müssen für eine Reliabilität von .701 mindestens 30 und in der Domäne Naturwissenschaft für eine Reliabilität von .705 ebenfalls 30 Items vorgelegt werden. Nach dem Abbruchkriterium von 48 Items wird für Lesen eine Reliabilität von .806, für Mathematik eine Reliabilität von .884 und für Naturwissenschaft eine Reliabilität von .813 erwartet. Die zu erwartende Reliabilität fallen beim FIT in allen Domänen geringer aus. In der Domäne Mathematik erreicht der FIT beispielsweise bei einer mittleren Länge von 17 Items (der spätere FIT sollte zwischen 32 und 36 Items lang sein) nur eine Reliabilität von .485 wohingegen der computerisierte adaptive Test bereits eine Reliabilität von .638 erreicht.

Für die CAT-Simulationsstudien wurde als Restriktion der MPI gewählt und so eingestellt, dass die Items pro Subdomäne gleichmäßig verteilt werden sollen. D. h., für die Domäne Lesen mit maximal 48 Items, sollen für jede der drei Subdimensionen 16 Items

vorgelegt werden; bei den Domänen Mathematik und Naturwissenschaft sind bei maximal 48 Items für jede der vier Subdimensionen 12 Items vorzulegen. Zur Kontrolle wurde bei den Simulationsergebnissen mit ausgegeben, wie viele Items jeder Subdimension vorgelegt wurden. Dabei lässt sich feststellen, dass die Content-Balancing-Methode MPI die Balancierung der Subdimensionen einwandfrei erfüllt.

4.4.3 Zusammenfassung

In diesem Kapitel wurden die wichtigsten Parameter des CAT-Algorithmus festgelegt. Zur Prüfung der Content-Balancing-Methode und zur Ermittlung der zu erwartenden Reliabilität wurde eine Simulationsstudie durchgeführt. Dabei bilden die selektierten Items aus der Kalibrierungsstudie und die Ergebnisse der Simulation die Grundlage für die Einstellungen der adaptiven Algorithmen in den drei Domänen. Der Personenparameterschätzer erhält zum Teststart einen Wert von 0 für jede Person. Als Personenparameterschätzer während der Testung wird der BME verwendet. Die Itemauswahl erfolgt anschließend nach der maximalen Information. Zu Beginn der Testung wird zufällig aus 10 passenden Items mit mittlerer Schwierigkeit ein Item ausgewählt. Der Test wird nach 48 Items bzw. 40 Minuten automatisch beendet. Als Restriktion wurde der MPI verwendet, welcher die Anteile der Items je Inhaltsbereich der betreffenden Domäne ausgleichen soll. Die Balancierung der Items pro Subdomäne erfolgte gleichmäßig. Die zu erreichende Messpräzision in Form der Reliabilität auf Grundlage der Simulationsstudie befindet sich bereits nach 30 Items je nach Domäne zwischen .7 und .8. Nach 48 Items erreichen alle drei Domänen eine Reliabilität von über .8. Konkret ist eine Reliabilität von mehr als .8 bei Lesen ab 21 Items, bei Mathematik ab 20 Items und bei Naturwissenschaft ab 28 Items zu erwarten. Die Reliabilität für den FIT fällt erwartungsgemäß geringer aus. Durch die Simulationsergebnisse ist deutlich geworden, dass CAT einen Vorteil gegenüber FIT u. a. dadurch bietet, dass eine gesteigerte Messpräzision und ein verringerter *SE* im Randbereich erzielt werden kann.

Anzumerken ist, dass die maximale Testlänge von 48 Items in der Pilotierungsstudie deutlich länger als die Testheftlänge in der Kalibrierungsstudie (33 Items) ist. Dies kann Auswirkungen auf die Gültigkeit der geschätzten Itemparameter haben. Beispielsweise kann der Positionseffekte ab der Position 40 deutlich zunehmen. Aufgrund der kurzen Testhefte der Kalibrierungsstudie geht dieses Wissen jedoch nicht in den Schwierig-

keitsparameter mit ein. Deshalb ist es ratsam, die Kalibrierungsstudie so zu planen, dass die Testhefte genauso lang wie die maximale Anzahl an vorzulegenden Items im adaptiven Test sind.

4.5 CAT – Veröffentlichung und Anwendung

In diesem Abschnitt werden die Pilotierungsstudie und deren Ablauf beschrieben. Ziele der Pilotierungsstudie waren u. a., die Funktionalität der adaptiven Tests technisch und psychometrisch zu prüfen. Aufgrund der empirischen Ergebnisse der Pilotierungsstudie konnte der adaptive Algorithmus angepasst und Empfehlungen für die Testanwendung in Form eines Manuals gegeben werden. Teilweise wurden nach der Pilotierungsstudie weitere Items entfernt und neue Simulationsstudien durchgeführt. Nach der Anwendung der entwickelten Tests in den weiteren Projekten der ASCOT-Initiative konnte die Bildung der endgültigen Skalen für die drei Domänen erfolgen und eine vorläufige Endversion des Tests je Domäne festgelegt werden. Hier wird der Begriff vorläufige Endversion verwendet, da es die Endversion für die ASCOT-Projekte war, so wie sie endgültig genutzt wurden. Wie weiter oben beschrieben, ist ein Test jedoch selten endgültig fertig und muss z. B. aufgrund von Parameterdrift stets angepasst werden. Neben einem computerisierten adaptiven Test wurde für jede Domäne auch ein FIT entwickelt und in der Pilotierungsstudie administriert. Nähere Informationen dazu befinden sich im Kapitel 4.6 Linking mit papierbasierter Testung.

4.5.1 Fragestellungen

- Wie ist der adaptive Algorithmus nach der Pilotierung anzupassen?
- Wie wird der Personenparameterschätzer zum Teststart spezifiziert?
- Wie erfolgt die Personenparameterschätzung während der Testung?
- Wie erfolgt die Itemauswahl zu Beginn der Testung?
- Wie erfolgt die Itemauswahl während der Testung?
- Nach welchen Kriterien wird der Test beendet?
- Erfüllt die Content-Balancing-Methode die Balancierung der Subdimensionen in diesem Itempool angemessen?

- Wie hoch ist die zu erzielende Messpräzision in den einzelnen Domänen bei der Nutzung des Itempools und der Verwendung des MPI auf Grundlage der empirischen Daten?
- Sind Items aus dem Itempool zu entfernen?
- Wie kann die Nachhaltigkeit des Tests sichergestellt werden?

4.5.2 Ablauf und Stichprobe: Pilotierungsstudie CAT

Bei der Pilotierungsstudie bekamen $N = 1\,093$ SuS einen computerisierten adaptiven Test genau einer der drei Domänen vorgelegt (Mathematik: $N = 390$ SuS; Lesen: $N = 350$ SuS; Naturwissenschaft: $N = 353$ SuS). Die Testhefte der einzelnen Domänen wurden innerhalb einer Klasse spiralisiert vorgegeben, so dass die Zuweisung der Domäne zufällig erfolgte. Im Mittel hat jeder Proband 35.507 Items ($SD = 12.897$ Items) bearbeitet. Die SuS waren durchschnittlich 22.064 Jahre ($SD = 3.735$ Jahre) alt. Die weiteren Häufigkeitsangaben zur Beschreibung der Stichprobe sind zur besseren Lesbarkeit als Stichpunkte dargestellt:

- Ausbildungsjahr: 0.7 % viertes Ausbildungsjahr; 71.0 % drittes Ausbildungsjahr; 20.3 % zweites Ausbildungsjahr; 7.3 % erstes Ausbildungsjahr; 0.6 % keine Angabe
- Geschlecht: 37.5 % weiblich; 62.1 % männlich; 0.4 % keine Angabe
- Schulabschluss: 29.8 % allgemeine Hochschulreife bzw. Fachhochschulreife; 61.4 % mittlere Reife; 7.2 % Haupt- bzw. Volksschulabschluss; 0.3 % ohne Schulabschluss; 0.2 % Abschluss der Polytechnischen Oberschule nach der 8. Klasse; 0.8 % Abschluss der Sonderschule bzw. Förderschule; 0.4 % keine Angabe
- Muttersprache: 86.0 % Deutsch; 12.4 % andere Sprache; 1.6 % keine Angabe
- Form der Berufsausbildung: 94.1 % duale Berufsausbildung; 5.7 % vollzeitschulische Berufsausbildung; 0.3 % keine Angabe
- Anzahl der Beschäftigten im Ausbildungsbetrieb: 18.7 % weniger als 10 Beschäftigte; 27.4 % zwischen 10 und 49 Beschäftigte; 21.2 % zwischen 50 und 249 Beschäftigte; 12.7 % zwischen 250 und 499 Beschäftigte; 15.9 % mit 500 und mehr Beschäftigten; 4.0 % keine Angabe oder in vollzeitschulischer Berufsausbildung
- Standort des Ausbildungsbetriebs: 33.9 % Hessen; 25.2 % Niedersachsen; 39.9 % Thüringen; 0.6 %anderes Bundesland; 0.5 % keine Angabe

- Berufsfeld: 15.8 % medizinisch/pflegender Bereich; 41.8 % gewerblich/technischer Bereich; 34.1 % kaufmännisch/verwaltender Bereich; 6.8 % anderes Berufsfeld; 1,5 % keine (plausible) Angabe
- Innerbetrieblicher Unterricht: 66.7 % innerbetrieblicher Unterricht; 32.7 % kein innerbetrieblicher Unterricht; 0.6 % keine Angabe

Nach der Schätzung der Personenfähigkeiten mittels der Software ConQuest ergab sich folgende Verteilung:

Tabelle 11

Fähigkeitsverteilung der Pilotierungsstichprobe für die Domänen Mathematik (MATH), Lesen (READ) und Naturwissenschaft (SCIE)

Domäne	θ_{mean}	$\sigma(\theta)^2$	θ_{min}	θ_{max}
READ	-0.115	0.846	-2.750	2.411
MATH	-0.084	0.839	-2.203	2.195
SCIE	-0.011	0.599	-2.576	1.891

Der Mittelwert der Fähigkeitsverteilung (θ_{mean}) liegt bei allen drei Domänen fast bei 0. Die Varianz ist bei den Domänen Mathematik und Lesen etwas kleiner als 1. In der Domäne Naturwissenschaft variieren die geschätzten Fähigkeiten der SuS hingegen weniger stark ($\sigma(\theta)^2 = 0.599$). Weiterhin bekamen 528 Probanden ein papierbasiertes Testheft mit fixer Itemreihenfolge vorgelegt. Eine detaillierte Stichprobenbeschreibung dazu befindet sich im Kapitel 4.6 Linking mit papierbasierter Testung.

4.5.3 Methode und Ergebnisse: Pilotierungsstudie CAT

Nach der Aufbereitung der Pilotierungsdaten wurden die Versionen der computerisierten adaptiven Tests und die dahinterliegenden adaptiven Algorithmen optimiert. Im Anschluss an dieses Vorgehen wurde für jede Domäne eine Endversion des computerisierten adaptiven Tests und des papierbasierten Tests mit fester Itemreihenfolge zur Nutzung in den ASCOT-Projekten erzeugt sowie ein ausführliches Nutzermanual (Bernhardt et al., 2013) geschrieben. Die computerisierten adaptiven Testversionen wurden den ASCOT-Projekten über einen Server zugänglich gemacht. Die Optimierung

der Algorithmen geschah u. a. dahingehend, optimale Abbruchkriterien zu finden, welche zu einer angemessenen zu erwartenden Reliabilität der Tests führen. Dazu wurden weiterführende Simulationen in einer zweiten Simulationsstudie durchgeführt. Die Simulationen erfolgten mit dem reduzierten Itempool, wie er nach der Pilotierungsstudie genutzt wurde (105 Mathematikitems, 94 Naturwissenschaftsitems und 65 Leseitems; vgl. Kapitel 4.3) in der Software MATE. Als Personenparameterschätzer wurde der BME gewählt. Die a-priori-Verteilung der Probanden für den BME ergab sich aus der Skalierung eines Rasch-Modells für die Daten aus den Kalibrierungsdaten mit den Items, wie sie am Ende für die Endversion in der ASCOT-Initiative verwendet worden. Konkret besaß die a-priori-Verteilung für alle drei Domänen den Mittelwert 0 und eine Varianz für Lesen von 0.739, für Mathematik von 0.954 und für Naturwissenschaft von 0.763. Die angenommene Verteilung der Probanden für die Itemauswahl war eine Standardnormalverteilung. Zu Beginn der Testung wurde für θ ein Wert von 0 für jede Person angenommen. Die Itemauswahl erfolgt nach der maximalen Iteminformation (vgl. Formel (14) auf S. 61). Zu Beginn wurde zufällig aus 10 passenden Items mit mittlerer Schwierigkeit ein Item gewählt. Die maximale Testlänge als mögliches Abbruchkriterium wurde für die Simulation so weit nach oben justiert, dass die maximale Itemanzahl des Itempools je Domäne bei gleichmäßiger Verteilung durch den MPI erreicht wird. Für den MPI ergaben sich dadurch folgende Einstellungen.

Tabelle 12

Content-Balancing-Restriktionen pro Inhaltsbereich für die Tests in den Domänen Lesen (READ), Mathematik (MATH) und Naturwissenschaft (SCIE) für die Simulationen der Endversionen

Domäne	Inhaltsbereich	Anzahl Items
READ	Deskriptionale Darbietung	18
	Gemischte Darbietung	18
	Depiktionale Darbietung	18
MATH	Quantität	23
	Veränderung und Beziehung	23
	Raum und Form	23
	Unsicherheit	23
SCIE	Leben und Gesundheit	20
	Erde, Planeten, Umwelt und natürliche Ressourcen	20
	Stoffe und Stoffveränderungen	20
	Bewegung, Kraft und Energie	20

Die Simulationen der zweiten Simulationsstudie wurden mit $N = 1\,000$ Personen durchgeführt. Die Ergebnisse der Simulation wurden darauf geprüft, (a) wie sich die zu erwartende Messpräzision (Reliabilität; vgl. Formel (22) auf S. 65) im Zusammenhang mit den Abbruchkriterien (Laufzeit und Itemanzahl) verhält und (b) ob die gleichmäßige Vorgabe von Items aller Subdomänen durch den MPI erfüllt wird. Anschließend wurden auch die empirischen Daten der Pilotierungsstudie auf die Punkte a und b hin geprüft. Bei der Pilotierungsstudie wurde die maximale Itemanzahl nicht immer erreicht. Auf Grundlage der Pilotierungsstudie wurde deshalb eine mittlere Bearbeitungszeit für ein Item je Domäne berechnet (Mathematik 64.9 Sek./Item, Lesen 100.0 Sek./Item und Naturwissenschaft 50.0 Sek./Item). In Bezug zur maximal möglichen Testzeit ergibt sich daraus eine mittlere Itemanzahl, die ein Proband in der maximal vorgegeben Testzeit im Mittel bearbeiten kann. Deshalb wurden die nachfolgenden Darstellungen um eine

sogenannte mittlere Reliabilität, welche auf der mittleren Itemanzahl beruht, ergänzt. Diese mittlere Reliabilität diente in der ASCOT-Haupterhebung auch dazu, dem Anwender eine Testlänge als Abbruchkriterium zu empfehlen. Die nachfolgenden drei Tabellen enthalten die Ergebnisse zu den Reliabilitätsberechnungen aufgrund der simulierten und der empirischen Daten. Konkret ist in Tabelle 13 (Domäne Lesen), Tabelle 14 (Domäne Mathematik) und Tabelle 15 (Domäne Naturwissenschaft) folgendes enthalten:

- Max. Zeit (in Sekunden und in Minuten): Nach Ablauf dieser Zeit wird der Test automatisch beendet.
- Max. Itemanzahl: Nach dem Abarbeiten der maximal möglichen Anzahl an Items wird der Test automatisch beendet.
- Max. Itemanzahl pro Inhaltsbereich: Zeigt die Verteilung der vorzulegenden Items auf die Inhaltsbereiche (Subdomänen) aus dem inhaltlichen Zielkonstrukt.
- Mittlere Itemanzahl: durchschnittliche Itemanzahl, die ein Proband in der maximal vorgegeben Testzeit bearbeiten kann.
- Geschätzte max. Reliabilität auf Grundlage der max. Itemanzahl und geschätzte mittlere Reliabilität auf Grundlage der mittleren Itemanzahl aus den Simulationsstudien (vgl. Formel (22) auf S. 65).
- Mittlere zu erwartende Reliabilität auf Grundlage der Pilotierungsstudie und der mittleren Itemanzahl (vgl. Formel (23) auf S. 65).

Für die Domäne Lesen wurden auf Grundlage der Ergebnisse für die Testlänge maximal 42 Items und für die Testzeit maximal 2100 sek. (35.00 min.) als Abbruchkriterium vorgeschlagen. In dieser Zeit wird eine mittlere Itemanzahl von 21 Items erwartet, was einer geschätzten mittleren Reliabilität von .8091 bzw. einer empirischen mittleren Reliabilität von .8073 entspricht (Bernhardt et al., 2013).

Tabelle 13

Reliabilität (Rel.) nach Abbruchkriterium für die Domäne Lesen

max. Zeit in Min.	max. Zeit in Sek.	max. Item-anzahl	max. Itemanzahl pro Inhalts-bereich	geschätz-te max. Rel.	mittlere Item-anzahl	geschätzte mittlere Rel.	empirische mittlere Rel.
2.50	150	3	1;1;1	.3678	1	.1646	.1733
5.00	300	6	2;2;2	.5499	3	.3678	.3847
7.50	450	9	3;3;3	.6491	4	.4456	.4542
10.00	600	12	4;4;4	.7106	6	.5499	.5530
12.50	750	15	5;5;5	.7545	7	.5902	.5902
15.00	900	18	6;6;6	.7865	9	.6491	.6488
17.50	1050	21	7;7;7	.8091	10	.6724	.6726
20.00	1200	24	8;8;8	.8293	12	.7106	.7110
22.50	1350	27	9;9;9	.8435	13	.7280	.7268
25.00	1500	30	10;10;10	.8552	15	.7545	.7529
27.50	1650	33	11;11;11	.8655	16	.7666	.7648
30.00	1800	36	12;12;12	.8742	18	.7865	.7840
32.50	1950	39	13;13;13	.8809	19	.7957	.7926
35.00	2100	42	14;14;14	.8873	21	.8091	.8073
37.50	2250	45	15;15;15	.8937	22	.8177	.8144
40.00	2400	48	16;16;16	.8978	24	.8293	.8255
42.50	2550	51	17;17;17	.9012	25	.8343	.8315
45.00	2700	54	18;18;18	.9045	27	.8435	.8412

Als Standardwert für die adaptiven Tests der Domäne Mathematik wurden 40 Items und 1297 sek. (21.62 min.) als Abbruchkriterium gewählt. Es wurde aufgrund der Ergebnisse davon ausgegangen, dass in der Zeit durchschnittlich mindestens 20 Items vorgelegt und beantwortet wurden. Bei dieser mittleren Itemanzahl ist eine mittlere Reliabilität von .8098 aufgrund der simulierten Daten und eine mittlere empirische Reliabilität von .8112 aufgrund der Daten der Pilotierungsstudie auszugehen. Beim Erreichen der maximalen Anzahl von 40 Items innerhalb der Zeit wird sogar eine Reliabilität von .8921 aufgrund der simulierten Daten erwartet (Bernhardt et al., 2013).

Tabelle 14

Reliabilität (Rel.) nach Abbruchkriterium für die Domäne Mathematik

max. Zeit in Min.	max. Zeit in Sek.	max. Item-anzahl	max. Itemanzahl pro Inhalts-bereich	geschätzte max. Rel.	mittlere Item-anzahl	geschätzte mittlere Rel.	empirische mittlere Rel.
2.16	130	4	1;1;1;1	.4553	2	.2998	.3031
4.32	259	8	2;2;2;2	.6303	4	.4553	.4633
6.49	389	12	3;3;3;3	.7175	6	.5603	.5638
8.65	519	16	4;4;4;4	.7727	8	.6303	.6331
10.81	649	20	5;5;5;5	.8098	10	.6816	.6831
12.97	778	24	6;6;6;6	.8354	12	.7175	.7212
15.14	908	28	7;7;7;7	.8546	14	.7512	.7510
17.30	1038	32	8;8;8;8	.8691	16	.7727	.7749
19.46	1168	36	9;9;9;9	.8814	18	.7942	.7949
21.62	1297	40	10;10;10;10	.8921	20	.8098	.8112
23.78	1427	44	11;11;11;11	.9001	22	.8249	.8257
25.95	1557	48	12;12;12;12	.9073	24	.8354	.8376

max. Zeit in Min.	max. Zeit in Sek.	max. Item-anzahl	max. Itemanzahl pro Inhalts-bereich	geschätzte max. Rel.	mittlere Item-anzahl	geschätzte mittlere Rel.	empirische mittlere Rel.
28.11	1686	52	13;13;13;13	.9132	26	.8463	.8482
30.27	1816	56	14;14;14;14	.9184	28	.8546	.8572
32.43	1946	60	15;15;15;15	.9224	30	.8632	.8652
34.59	2076	64	16;16;16;16	.9258	32	.8691	.8716
36.76	2205	68	17;17;17;17	.9289	34	.8762	.8783
38.92	2335	72	18;18;18;18	.9317	36	.8814	.8835
41.08	2465	76	19;19;19;19	.9343	38	.8880	.8887
43.24	2595	80	20;20;20;20	.9368	40	.8921	.8927
45.41	2724	84	21;21;21;21	.9389	42	.8969	.8968
47.57	2854	88	22;22;22;22	.9405	44	.9001	.9002
49.73	2984	92	23;23;23;23	.9415	46	.9042	.9034

Für die Domäne Naturwissenschaft wurden für die Testlänge maximal 56 Items und für die Testzeit maximal 1400 sek. (23.33 min.) als Abbruchkriterium vorgeschlagen. In dieser Zeit wird eine mittlere Itemanzahl von 28 Items erwartet, was einer geschätzten mittleren Reliabilität von .8478 bzw. einer empirischen mittleren Reliabilität von .8035 entspricht (Bernhardt et al., 2013).

Tabelle 15

Reliabilität (Rel.) nach Abbruchkriterium für die Domäne Naturwissenschaft

max. Zeit in Min.	max. Zeit in Sek.	max. Item-anzahl	max. Itemanzahl pro Inhalts-bereich	geschätzte max. Rel.	mittlere Item-anzahl	geschätzte mittlere Rel.	empirische mittlere Rel.
1.67	100	4	1;1;1;1	.4491	2	.2929	.2328
3.33	200	8	2;2;2;2	.6135	4	.4491	.3777
5.00	300	12	3;3;3;3	.7093	6	.5479	.4759
6.67	400	16	4;4;4;4	.7611	8	.6135	.5467
8.33	500	20	5;5;5;5	.7956	10	.6689	.6016
10.00	600	24	6;6;6;6	.8228	12	.7093	.6429
11.67	700	28	7;7;7;7	.8478	14	.7405	.6776
13.33	800	32	8;8;8;8	.8642	16	.7611	.7053
15.00	900	36	9;9;9;9	.8752	18	.7809	.7291
16.67	1000	40	10;10;10;10	.8848	20	.7956	.7486
18.33	1100	44	11;11;11;11	.8919	22	.8114	.7661
20.00	1200	48	12;12;12;12	.8986	24	.8228	.7797
21.67	1300	52	13;13;13;13	.9047	26	.8369	.7931
23.33	1400	56	14;14;14;14	.9110	28	.8478	.8035
25.00	1500	60	15;15;15;15	.9165	30	.8570	.8143
26.67	1600	64	16;16;16;16	.9203	32	.8642	.8226
28.33	1700	68	17;17;17;17	.9238	34	.8705	.8313
30.00	1800	72	18;18;18;18	.9269	36	.8752	.8375

max. Zeit in Min.	max. Zeit in Sek.	max. Item-anzahl	max. Itemanzahl pro Inhalts-bereich	geschätzte max. Rel.	mittlere Item-anzahl	geschätzte mittlere Rel.	empirische mittlere Rel.
31.67	1900	76	19;19;19;19	.9287	38	.8803	.8445
33.33	2000	80	20;20;20;20	.9304	40	.8848	.8496

Um den Verlauf der Reliabilität besser einschätzen zu können, wurde in den nachfolgenden Abbildungen der Verlauf der (a) geschätzten maximalen Reliabilität, (b) der geschätzten mittleren Reliabilität und (c) der empirischen mittleren Reliabilität in Anhängigkeit von der Anzahl vorgelegter Items für die Domänen Lesen, Mathematik und Naturwissenschaft abgebildet.

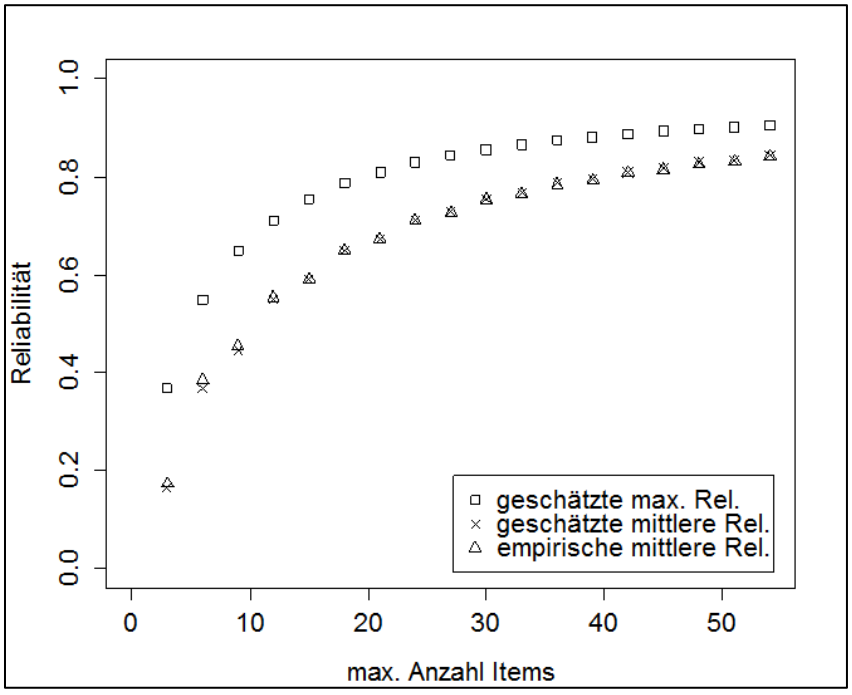

Abbildung 18: Zu erwartende Reliabilität in Abhängigkeit der Testlänge für die Domäne Lesen.

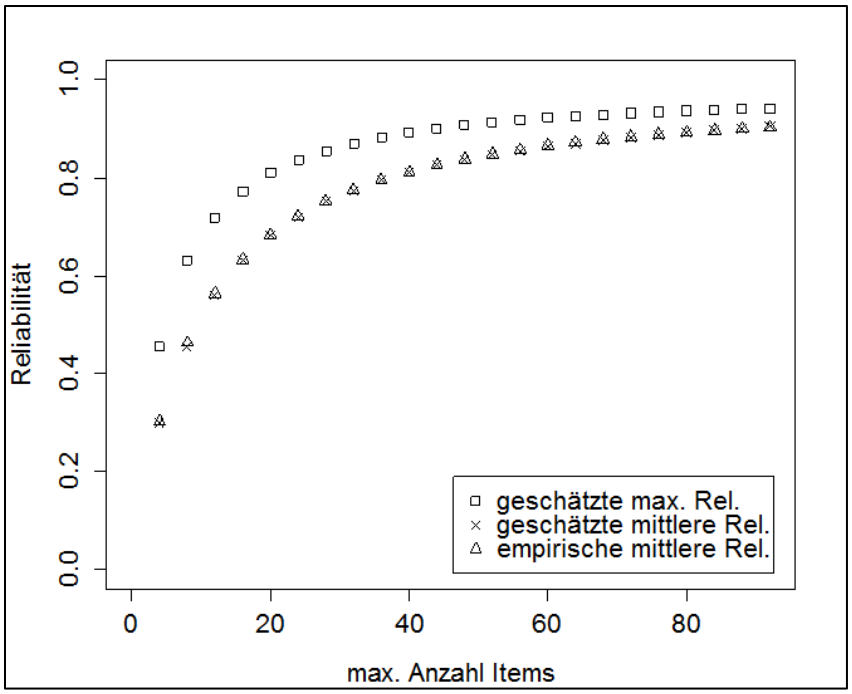

Abbildung 19. Zu erwartende Reliabilität in Abhängigkeit der Testlänge für die Domäne Mathematik.

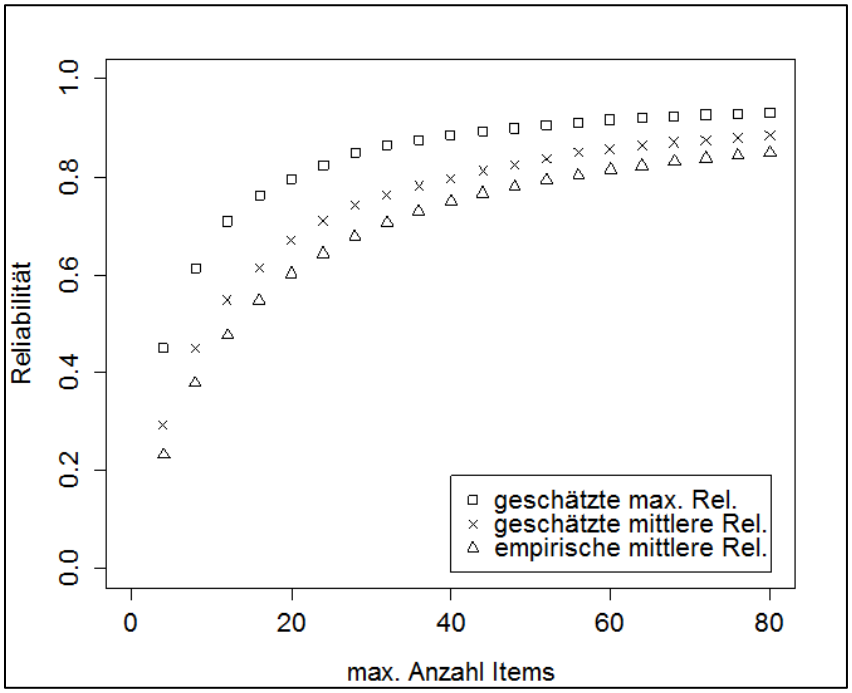

Abbildung 20. Zu erwartende Reliabilität in Abhängigkeit der Testlänge für die Domäne Naturwissenschaft.

In den Domänen Mathematik und Lesen sind die mittleren Reliabilitäten für die Daten der Simulationsstudie und die Daten der Pilotierungsstudie fast identisch. In der

Domäne Naturwissenschaft fällt die empirische mittlere Reliabilität im Vergleich zur geschätzten mittleren Reliabilität etwas geringer aus. Auf Grundlage der Ergebnisse der Pilotierungsstudie und der zweiten Simulationsstudie wurden keine weiteren Parameter des Algorithmus verändert. D. h., als Personenparameterschätzer wurde weiterhin der BME mit den oben angegebenen a-priori-Informationen gewählt. Zu Beginn wurde für θ ein Wert von 0 für jede Person angenommen. Die Itemauswahl erfolgt nach der maximalen Information. Zu Beginn wurde zufällig aus 10 passenden Items mit mittlerer Schwierigkeit ein Item gewählt. Als Restriktion wurde der MPI verwendet, welcher die Anteile der Items je Inhaltsbereich der betreffenden Domäne ausgleichen soll. Die Balancierung der Items pro Subdomäne erfolgte nach einer Prüfung der empirischen Daten gleichmäßig. Da der Test in der Realität selten wie bei einer Simulation genau nach einer gleichverteilten Vorgabe der Items pro Subdimension abbricht, wurde der Test in der Pilotierungsstudie häufig beendet, bevor eine Person in allen Subdomänen die gleiche Anzahl an Items vorgelegt bekommen hat. In der Domäne Mathematik wurden beispielsweise nach Erreichen der maximalen Testzeit nur 18 Items vorgelegt. Der MPI glich in der Pilotierungsstudie aus, dass die Unterschiede zwischen der Anzahl vorgegebener Items zwischen den Subdimensionen maximal 1 ist. Die Verteilung könnte über die vier Subdimensionen dann folgendermaßen aussehen: 5 Items in der ersten Subdimension, 4 Items in der zweiten Subdimension, 5 Items in der dritten Subdimension und 4 Items in der vierten Subdimension. Über die gesamte Personenzahl hinweg ist die Verteilung der vorgelegten Items deshalb nicht in allen Subdomänen hundertprozentig gleich. Die genaue Verteilung der Erfüllung des MPI durch den Algorithmus für die Daten der Pilotierungsstudie sind in der Tabelle 16 abzulesen.

Tabelle 16

Relativer Anteil an vorgegebenen Items pro Subdomäne für die Domänen Mathematik (MATH), Lesen (READ) und Naturwissenschaft (SCIE).

	prozentualer Anteil vorgelegter Items in			
Domäne	Subdomäne 1	Subdomäne 2	Subdomäne 3	Subdomäne 4
READ	33.630	33.812	32.558	-
MATH	24.922	25.343	25.271	24.464
SCIE	24.939	25.081	25.171	24.809

Anmerkung: Subdomäne 4 entfällt bei Lesen, da in dieser Domäne nur drei Subdomänen verwendet wurden.

4.5.4 Methode: Wartung und Pflege

Um die Nachhaltigkeit eines computerisierten adaptiven Tests sicherzustellen, sind, wie im Kapitel 3.6 theoretisch beschrieben, die Erhaltung der Skala und somit die Wartung der Itempools und der adaptiven Algorithmen der drei Tests über die verwendete Software MATE notwendig. Die Arbeit an den Tests endet weder mit der Pilotierungsstudie noch mit der Erstellung der vorläufigen Endversionen für die ASCOT-Initiative und auch nicht mit der Weitergabe der Tests an ein Datenzentrum. Damit die Tests in den drei Domänen auch in Zukunft korrekt funktionieren, ist es von Bedeutung, Wartungsintervalle einzuführen, um den Itempool (Entfernen und Hinzufügen von Items sowie Prüfung des Itemparameterdrift) und den adaptiven Algorithmus (Anpassung von Abbruchkriterien, Constraints wie MPI usw. an den Itempool) zu pflegen und ggf. anzupassen. Zudem sollte aus technischer Sicht die Software gewartet und angepasst bzw. die Tests ggf. in andere Softwarelösungen implementiert werden. D. h., da die Pflege und Ausführung der Tests über die Software MATE erfolgt, muss auch diese aktuell gehalten werden. Neue Anforderungen, wie z. B. neue Auslieferungsmodi, können sonst nicht erfüllt und die Kompatibilität mit aktuellen technischen Gegebenheiten, wie z. B. neue Betriebssysteme, nicht mehr gewährleistet werden. Für das Projekt MaK-adapt ist im Projektantrag dazu keine Lösung vorgesehen. Der Stand am Ende des Projektes ist, dass für jede Domäne ein adaptiver Test mit jeweils fünf unterschiedlichen Testlängen je

Domäne als Offline-Version zur Verfügung steht (Bernhardt, Frey, Ziegler & Seeber, 2016). Um den Test nutzen zu können, muss ein Ordner mit den Tests und der Software auf den zu nutzenden Computer übertragen werden. Über eine lokale Datei wird der Test anschließend gestartet. Wartungen im Sinne der Überprüfung des Itempools und deren Parameter, einer Anpassung des adaptiven Algorithmus oder einer softwaretechnischen Änderung sind nicht vorgesehen. Aus diesem Grund wird an dieser Stelle nur eine mögliche Vorgehensweise vorgeschlagen, die über das Projekt MaK-adapt hinausgeht und empirisch nicht geprüft wurde.

An erster Stelle sollte eine regelmäßige Routine ausgeführt werden, mit welcher die Gültigkeit der Itemparameter (hier konkret die Prüfung der Itemschwierigkeiten) und das Itemmaterial auf ihre Qualität hin (z. B. Itemfit, Trennschärfe, Aktualität und Gültigkeit der Iteminhalte für die zu prüfende Stichprobe, DIF-Prüfung auf aktuelle Berufe und weitere relevante Kovariate) geprüft werden (vgl. Kapitel 4.3.4). Die Prüfung kann auf der Grundlage von Testdaten erfolgen, welche durch die Nutzung der Tests durch Dritte anfallen. Dafür müssen vor der Nutzung der Tests durch Dritte entsprechende Kooperationsvereinbarungen getroffen werden. Es sind aber auch spezielle Wartungsstudien denkbar, in denen eigenständige Erhebungen mit den Tests durchgeführt werden, ausschließlich um die Qualität der Itempools und Algorithmen zu testen. Da solche Studien einen Kostenfaktor darstellen, welcher nur schwer zu decken ist, wären sie im konkreten Fall im Rahmen von Abschlussarbeiten oder Forschungsseminaren möglich, in welchen Studierende anhand einer selbst rekrutierten Stichprobe die Tests durchführen und die Ergebnisse auswerten. Dies wiederum ist nur bei höheren Mastersemestern vorstellbar, welche bereits Erfahrungen mit IRT und Testentwicklung sammeln konnten.

Die Prüfung der Qualität der Itempools in festen Intervallen führt höchstwahrscheinlich zu einem Ausschluss verschiedener Items über die Zeit hinweg. Daher ist es notwendig, regelmäßig neue Items zu produzieren bzw. entsprechend der vorgeschlagenen Methoden zur Itemwiederverwertung zu adaptieren (vgl. Kapitel 4.2.2), zu kalibrieren und mit den Itemschwierigkeiten der vorhandenen Skala zu verbinden (Linking). Hier werden separate Kalibrierungsstudien empfohlen. Denn das Mitlaufen neuer Items in einem adaptiven Test und die anschließende Kalibrierung können zu Problemen führen. Z. B. erhalten Probanden mit neuen Items bei gleicher Reliabilität einen längeren

Test (vgl. Kapitel 3.6.3). Zudem kann bei einer separaten Kalibrierungsstudie ein fixes Testheftdesign ähnlich wie bei der ursprünglichen Kalibrierungsstudie verwendet werden (vgl. Kapitel 4.3). Dies ermöglicht eine hohe Vergleichbarkeit der Testbedingungen, was für das Linking eine gute Voraussetzung ist (vgl. Kapitel 3.7). Weiterhin ist bei einem entsprechenden Testheftdesign die Betrachtung der Itempositionseffekte nach wie vor möglich (vgl. Kapitel 4.3.5). Bei der Kalibrierung der neuen Items können eine Auswahl alter Items oder alle alten Items mitlaufen. Diese können später als Ankeritems genutzt werden. So kann (a) ermittelt werden, ob sich die alten Items vergleichbar zur ursprünglichen Kalibrierung verhalten (z. B. gleiche Itemschwierigkeiten bei gleicher Stichprobe) und (b) können die neuen Items mit der vorhandenen Skala des adaptiven Tests verbunden werden (vgl. Linking mit Ankeritems z. B. Kapitel 4.6.3).

In Bezug auf den adaptiven Algorithmus können die Tests regelmäßig daraufhin geprüft werden, ob in der vorgegebenen max. Zeit bzw. nach der vorgegebenen max. Anzahl an Items noch die zu erwartende Reliabilität erreicht wird. Bei einer größeren Menge neu hinzugefügter Items kann dies vorab über Simulationsstudien geprüft werden. Zudem könnten die Testanwender in ihren Studien die empirische Reliabilität ermitteln und nach den Testungen an den Testentwickler rückmelden. Die Fragen, wer wann (a) diese Wartungsarbeiten durchführt, (b) diese Änderungen über die Software MATE einpflegt, (c) die Lauffähigkeit der Software sicherstellt und ggf. (d) die Erweiterungen der Anwendung durch Anpassung der Software bzw. die Nutzung anderer Software sicherstellt, stellten sich als Herausforderung dar, die am Projektende nicht geklärt waren. Damit die Tests lange als qualitativ hochwertiges Instrument genutzt werden können, sind diese Fragen zu beantworten. Für die Entwicklung von Tests wird deshalb an dieser Stelle empfohlen, solche Überlegungen bereits in die Testplanung mit einfließen zu lassen.

4.5.5 Zusammenfassung

Nach der Pilotierungsstudie wurde den ASCOT-Projekten die Möglichkeit gegeben, sich aus den drei adaptiven Tests und den drei papierbasierten Testheften für jede Domäne, ein oder mehrere Tests auszusuchen. Jedes ASCOT-Projekt konnte selbstständig entscheiden, welche Reliabilität angestrebt wird und so die Testlänge auf Grundlage der Tabelle 13, Tabelle 14 und Tabelle 15 für die computerisierten adaptiven Tests

bestimmen. Die Reliabilitätsanalysen beruhen sowohl auf den empirischen Daten der Pilotierungsstudie als auch auf neu simulierten Daten. Bis auf die individuelle Testzeit und Testlänge waren die Parameter des adaptiven Algorithmus für alle drei Domänen gleich und wurden bis auf die a-priori-Information für den BME im Vergleich zur Festlegung für die Pilotierungsstudie nicht geändert. Die adaptiven Tests für die Nutzung in den ASCOT-Projekten wurden über einen Server in Jena online bereitgestellt. Mit der Beendigung der Projektlaufzeit endete auch die Bereitstellung der Tests als Onlineversion, da diese zusätzliche Ressourcen für den Erhalt und die Pflege der Tests benötigt. Nach Projektende stand eine Offlineversion der adaptiven Tests zur Verfügung (Bernhardt et al., 2016). Zusätzlich ergaben sich nach Abschluss des gesamten Projektes noch einmal minimale Änderungen am Itempool, da nicht für alle Items die Berechtigungen vorlagen, diese auch außerhalb der ASCOT-Initiative zu nutzen. Der Itempool für die Domäne Mathematik enthält abschließend 102 von 105 Items und der Itempool für die Domäne Lesen 62 von 65 Items. In der Domäne Naturwissenschaft konnten alle 94 Items beibehalten werden. Für die papierbasierten Testhefte gab es für jede Domäne in der Pilotierungsstudie genau eine Version. Diese papierbasierte Version kann nach der Dateneingabe, -aufbereitung und -auswertung anschließend mit der Skala des computerisierten adaptiven Tests verbunden werden. Nähere Angaben dazu befinden sich im Kapitel 4.6. Die Ergebnisse der Pilotierungsstudie entsprechen denen der Simulationsstudien und zeigen, dass auf Basis der Itempools eine hinreichende Messpräzision mit geringer Itemanzahl beim adaptiven Testen erzielt werden kann. Zudem kann der MPI als zuverlässiges Instrument des Content-Balancing beim computerisierten adaptiven Testen empfohlen werden, wenn der Itempool relativ gleichverteilt angelegt wird.

4.6 Linking mit papierbasierter Testung

In Studien können unterschiedliche Testformen eingesetzt werden (z. B. papierbasiertes vs. computerisiertes Testen, konventionelles Testen mit fester Itemreihenfolge vs. adaptives Testen). Die Testergebnisse unterschiedlicher Testformen sollen am Ende häufig auf einer gemeinsamen Metrik berichtet werden. Zu diesem Zweck können Linkprozeduren verwendet werden. Ein Linking setzt aber u. a. invariante Itemparameter über verschiedene Testformen voraus. Aufgrund konstruktirrelevanter Faktoren, wie z. B. Änderungen in Itemposition, Text, Testzeit, Design, Bedingungen, usw., können

Itemparameter zwischen Testformen variieren (Kolen & Brennan, 2014; Miller & Fitzpatrick, 2008). Itempositionseffekte oder die Art des verwendeten Testheftdesigns als Ursachen für die Variation der Itemparameter zwischen Testformen werden jedoch selten berücksichtigt. Aktuelle Studien legen nahe, dass bei der Schätzung von Itemparametern Positionseffekte zu berücksichtigen sind (Albano, 2013; Debeer & Janssen, 2013; Hartig & Buchholz, 2012). Hier wird deshalb ein Ansatz zur Berücksichtigung von Itempositionseffekten beim Linking und dessen Auswirkungen an einem empirischen Beispiel gezeigt. Konkret wird im letzten Abschnitt dieses Kapitels beschrieben, wie die Skala eines papierbasierten Tests mit fester Itemreihenfolge und die Skala eines computerisierten adaptiven Tests verbunden werden können. Dazu werden die Erkenntnisse zu den Itempositionseffekten aus der Kalibrierungsstudie genutzt. Die vorgestellte Methode wird zudem empirisch geprüft.

4.6.1 Fragestellungen

- Wie lassen sich die Itempools eines computerisierten adaptiven Tests und eines papierbasierten Tests mit fester Itemreihenfolge angemessen miteinander verbinden?
- Welche Auswirkungen hat die Verwendung der Linkprozedur mit Beachtung von Itempositionseffekten auf die Auswahl der Linkitems?
- Welche Auswirkungen hat die Linkingprozedur mit Beachtung von Itempositionseffekten auf die Personenparameterverteilung und die Reliabilität des FIT?

4.6.2 Ablauf und Stichprobe: Pilotierungsstudie papierbasierte Testung

Bei der Pilotierungsstudie bekamen 528 SuS einen papierbasierten Test mit genau einer Domäne mit fester Itemreihenfolge vorgelegt (179 Personen Lesen, 176 Personen Mathematik und 173 Personen Naturwissenschaft). Im Mittel wurden 34.049 Items (*SD* = 6.111 Items) pro Person bearbeitet. Die SuS waren im Durchschnitt 23.465 Jahre (*SD* = 6.354 Jahre) alt. Die weiteren Häufigkeitsangaben zur Beschreibung der Stichprobe sind zur besseren Lesbarkeit als Stichpunkte dargestellt:

- Ausbildungsjahr: 0.6 % viertes Ausbildungsjahr; 70.5 % drittes Ausbildungsjahr; 26.3 % zweites Ausbildungsjahr; 1.7 % erstes Ausbildungsjahr; 0.9 % keine Angabe

- Geschlecht: 38.4 % weiblich; 61.0 % männlich; 0.6 % keine Angabe
- Schulabschluss: 19.5 % allgemeine Hochschulreife bzw. Fachhochschulreife; 66.5 % mittlere Reife; 11.9 % Haupt- bzw. Volksschulabschluss; 0.8 % ohne Schulabschluss, Abschluss der Polytechnischen Oberschule nach der 8. Klasse oder Abschluss der Sonderschule bzw. Förderschule; 1,3 % keine Angabe
- Muttersprache: 86.0 % Deutsch; 10.0 % andere Sprache; 4.0 % keine Angabe
- Form der Berufsausbildung: 87.8 % duale Berufsausbildung; 10.6 % vollzeitschulische Berufsausbildung; 1.7 % keine Angabe
- Anzahl der Beschäftigten im Ausbildungsbetrieb: 15.5 % weniger als 10 Beschäftigte; 24.6 % zwischen 10 und 49 Beschäftigte; 23.9 % zwischen 50 und 249 Beschäftigte; 10.2 % zwischen 250 und 499 Beschäftigte; 18.2 % mit 500 und mehr Beschäftigten; 7.6 % keine Angabe oder in vollzeitschulischer Berufsausbildung
- Standort des Ausbildungsbetriebs: 40.7 % Hessen; 18.6 % Niedersachsen; 30.5 % Thüringen; 8.7 % anderes Bundesland; 1.5 % keine Angabe
- Berufsfeld: 25.6 % medizinisch/pflegender Bereich; 42.0 % gewerblich/technischer Bereich; 22.0 % kaufmännisch/verwaltender Bereich; 5.9 % anderes Berufsfeld; 1,5 % keine (plausible) Angabe
- Innerbetrieblicher Unterricht: 66.1 % innerbetrieblicher Unterricht; 32.6 % kein innerbetrieblicher Unterricht; 1.3 % keine Angabe

Die Items der Testhefte im papierbasierten Test wurden zum Großteil aus dem Itempool der computerisierten adaptiven Tests entnommen. Es gab 33 Leseitems (31 Items aus dem adaptiven Test und zwei neue Items), 36 Mathematikitems (alle 36 Items aus dem adaptiven Test) und 41 Naturwissenschaftsitems (36 Items aus dem adaptiven Test und fünf neue Items) im Testheft der jeweiligen Domäne. Es wurden teilweise neue Items in den Testheften mit dem Ziel untergebracht, diese später auch in dem adaptiven Test einzubringen. Auf Grund der nachfolgend vorgestellten Linking-Methode ist es theoretisch sogar möglich, die kalibrierten Itemkennwerte für diese Items auf die Metrik des adaptiven Tests umzuwandeln und sie dort einzusetzen. Praktisch wurde dieses Vorgehen jedoch nicht umgesetzt. Die Testhefte der einzelnen Domänen wurden innerhalb einer Klasse spiralisiert vorgegeben, so dass die Zuweisung der Domäne

zufällig erfolgte. Um keine systematische Variation der Itemparameter zwischen den beiden Testformen durch konstruktirrelevante Faktoren hervorzurufen, wurden die Item-Folien aus dem computerisierten adaptiven Test identisch auf das Papier übertragen. Dabei wurde darauf geachtet, dass jedes mehrseitige Item einzeln steht, also nicht zwei Items auf einer Seite vorhanden sind. Durch das identische Abbilden der Item-Folien auf Papier wird sichergestellt, dass Textänderung, Änderung der Antwortreihenfolge, Änderungen im Design der Aufgabenblätter (Schreibstil, Aufgabenstellung, Schrift etc.) möglichst keine Rolle spielen. Zudem wurden die Testbedingungen gleich gehalten, was z. B. das Timing (die Testzeit), die motivationalen Bedingungen, die Hilfsmittel wie Schmierpapier und Taschenrechner oder die Bedingungen der Testräume (Schule) anbelangt. Bei Kahlecke (2014) wurde für die beiden vorliegenden Testversionen (papierbasiertes FIT und CAT) mittels der *Student Opinion Scale* geprüft, ob ein Unterschied bei der Leistungsmotivation zur Testbearbeitung zwischen den Testteilnehmern besteht. Hier wurden keine Unterschiede in der Motivation gefunden, was eine wichtige Voraussetzung dafür ist, die beiden Testversionen als gleichwertig anzusehen und ihre Skalen entsprechend miteinander verbinden zu können. Ein Bekanntwerden der Iteminhalte ist ebenfalls auszuschließen, da der Testzeitraum in derselben Zeit lag, in der auch die Pilotierung des computerisierten adaptiven Tests stattfand. Bei der Instruktion wurde darauf hingewiesen, dass die Teilnehmer alle Items nacheinander beantworten sollen und kein Item auslassen dürfen. Im Zweifelsfall wurden sie aufgefordert, die Antwort zu raten. Es wurde hierbei besonders durch die geschulten Testleiter sichergestellt, dass nicht vor- und zurückgeblättert wurde. Dadurch kann die Annahme getroffen werden, dass gleiche Positionseffekte wie beim computerisierten adaptiven Test vorliegen, wo ein Item-Review technisch unterbunden wird. Nachfolgend wird die Instruktion des papierbasierten Tests wörtlich wiedergegeben.

> Liebe Teilnehmerin, lieber Teilnehmer,
>
> vielen Dank für Ihre Bereitschaft an unserer Studie teilzunehmen. Bei dieser werden Aufgaben zur Messung der Kompetenzen von Berufsschülerinnen und Berufsschülern in den Bereichen Mathematik, Lesen und Naturwissenschaften erprobt. Die Tests werden später deutschlandweit an Berufsschulen zur Kompetenzmessung eingesetzt werden.

Die Teilnahme an der Studie ist freiwillig. Ihre Angaben sind nur Mitarbeiterinnen und Mitarbeitern des Forschungsprojekts „Messung allgemeiner Kompetenzen – adaptiv“ zugänglich, werden ohne Namen gespeichert und nicht an Ihre Schule zurückgemeldet. Die Auswertung der Daten erfolgt anonymisiert. Leistungen einzelner Personen werden nicht ausgewertet. Die Ergebnisse dienen ausschließlich wissenschaftlichen Zwecken.

Die Untersuchung wird insgesamt ca. 90 Minuten dauern. Zu Beginn werden wir Ihnen einige Fragen zu Ihrer Person stellen. Bitte beantworten Sie diese wahrheitsgemäß.

In den darauffolgenden 40 Minuten bekommen Sie Aufgaben aus den Bereichen Mathematik oder Lesen oder Naturwissenschaften vorgelegt. Bitte lesen Sie sich die Aufgabenstellung genau durch und klicken Sie danach die Antwort an, die Ihrer Meinung nach richtig ist. Es ist jeweils genau eine Antwort richtig. Bei einigen Aufgaben sind auch Zahlen oder einzelne Wörter einzutragen.

Anschließend werden Ihnen einige weitere Fragen gestellt, die der Beurteilung der Tests und der Testbearbeitung dienen sollen.

Es ist jeweils genau eine Antwort richtig. Bitte kreuzen Sie NUR EINE Antwortmöglichkeit an – Mehrfachantworten sind nicht möglich.

Wichtig zu wissen ist, dass die Aufgaben unterschiedliche Schwierigkeitsgrade haben. Es wird somit vorkommen, dass einige Aufgaben von Ihnen als eher leicht und andere als eher schwer empfunden werden.

Für den Erfolg der Studie ist es wichtig, dass Sie alle Aufgaben konzentriert und so gut wie möglich bearbeiten. Sollten Sie eine Aufgabe einmal nicht sicher lösen können, dann geben Sie bitte die Antwort an, die Ihrer Meinung nach am ehesten stimmt.

Einige Aufgaben erstrecken sich über mehrere Seiten. Bitte achten Sie beim Bearbeiten darauf, alle Seiten zu berücksichtigen.

Anschließend werden Ihnen weitere Fragen gestellt, die der Beurteilung der Tests und der Testbearbeitung dienen sollen.

Sollten Sie noch Fragen zum Testablauf haben, dann können Sie sich an die Testleiterin bzw. den Testleiter wenden. Dieser wird, sobald alle fertig mit dem Lesen sind, eine entsprechende Frage stellen.

Vielen Dank für Ihre Teilnahme und viel Erfolg!

Beim computerisierten adaptiven Test war es im Verlauf des Tests nicht möglich weiterzugehen, ohne zuvor eine Antwort zu geben. Bei der papierbasierten Testung bestand die Möglichkeit, Aufgaben zu überspringen. Dies kann fehlende Antworten verursachen. In der Instruktion wurde deshalb der Hinweis gegeben, jedes Item zu beantworten, der Reihe nach das Testheft zu bearbeiten und ggf. zu raten. Es liegt somit die Annahme zugrunde, dass die Aufgaben nacheinander durchgeblättert wurden und jedes Item zumindest angeschaut wurde. Ein nichtbeantwortetes Item, innerhalb eines Antwortblockes, gilt daher als falsch. Fehlende Werte am Ende der Testung galten nach wie vor als fehlender Wert.

4.6.3 Methode und Ergebnisse: Linking

Aufgrund der Angaben zum Testablauf und zur Stichprobe kann die Annahme getroffen werden, dass ein Gruppendesign mit nicht äquivalenten Gruppen mit der Verwendung gemeinsamer Items (Kolen & Brennan, 2014) bei der Nutzung unterschiedlicher Testformen mit unterschiedlichen Testheftdesigns vorliegt. Wie im theoretischen Teil (vgl. Kapitel 3.7.2) festgestellt, muss bei diesem Datenerhebungsdesign jedoch sichergestellt werden, dass die gemeinsamen Items in derselben Reihenfolge in den unterschiedlichen zu verbindenden Testformen vorkommen. Da es beim papierbasierten Testheft genau eine Version gab und bei der Kalibrierung der Items für CAT viele unterschiedliche Testhefte verwendet wurden, wird an dieser Stelle das Wissen um die Positionseffekte verwendet, um die Gleichheit der Itemreihenfolge zu simulieren. Konkret wurde die Linkingprozedur, wie im nachfolgenden Flussdiagramm beschrieben, angewandt.

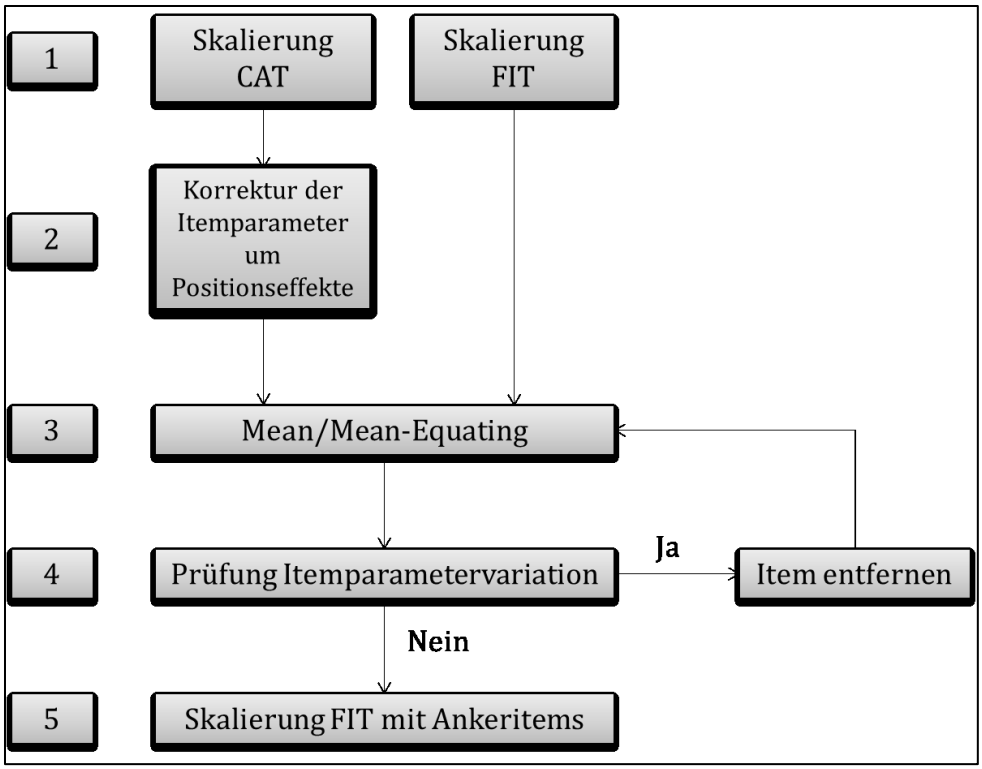

Abbildung 21. Flussdiagramm für die Linkingprozedur.

Schritt 1: Skalierung CAT und FIT

Im ersten Schritt wurden die Items der beiden Testformen (CAT und papierbasiertes FIT) mit einem Rasch-Modell mithilfe der Software ConQuest 3.0.1 (Adams et al., 2012) frei skaliert. Die Verteilung der geschätzten Personenfähigkeit θ wurde im Mittel auf 0 fixiert. Die Syntax zur Fixierung der Personenfähigkeit und zur Schätzung eines einfachen Rasch-Modells für die Itemparameter in der Software ConQuest lautet gleichbleibend zur Syntax der Kalibrierung (vgl. Kapitel 4.3.4):

```
set constraints=cases;

model Item;
```

Da es sich bei den beiden Tests um ein nicht-äquivalentes Gruppendesign mit der Verwendung gemeinsamer Items handelt, müssen die sogenannten Linkitems bzw. Ankeritems ermittelt werden. Hier wurden alle gemeinsamen Items beider Testformen im ersten Schritt als Linkitems bzw. Ankeritems gewählt.

Schritt 2: Korrektur der Itemparameter um Positionseffekte

Aus den vorherigen Studien zu Positionseffekten (vgl. Kapitel 4.3.5) ist bekannt, dass die Itemparameter aus der Kalibrierung für die computerisierten adaptiven Tests die durchschnittlichen Itempositionseffekte beinhalten. Hier wird die Annahme zugrunde gelegt, dass der papierbasierte Test mit fester Itemreihenfolge ebenfalls Positionseffekte enthält, die aufgrund eines nicht vorhandenen Testheftdesigns jedoch nicht ermittelt werden können. Deshalb wird die Itemschwierigkeit der Items aus dem papierbasierten Test nicht mit der mittleren Itemschwierigkeit aus dem adaptiven Test verglichen, sondern das Wissen über die Itempositionseffekte aus dem computerisierten adaptiven Test ($\gamma_{k_{CAT}}$) wird hinzugezogen. Konkret heißt das, dass die Itemschwierigkeiten der Items im adaptiven Test (b_{CAT}) um den Itempositionseffekt ergänzt (summiert) werden. Somit erhält man die Itemschwierigkeit für ein Item an Position k im adaptiven Test und kann es mit dem entsprechenden Item im papierbasierten Test an der Position k vergleichen.

Schritt 3: Mean/Mean-Equating

Im dritten Schritt wurden die unterschiedlichen Verteilungen der Schwierigkeitsparameter auf einen einheitlichen Mittelwert verschoben bzw. die unterschiedliche Verteilung der Itemschwierigkeiten des papierbasierten Tests (b_{FIT}) zu b_{CAT} durch eine lineare Transformation mittels der Mean/Mean-Methode (Loyd & Hoover, 1980) auf die Metrik des computerisierten adaptiven Tests gebracht. D. h., die Itemschwierigkeiten der Items aus dem papierbasierten Test (b_{FIT}) wurden neu berechnet zu $b_{FIT_{M/M}}$, und zwar durch die Summe der Itemschwierigkeiten aus dem papierbasierten Test und einer Verschiebung, dem sogenannten Shift (Shift+b_{FIT}). Der Shift ergab sich aus der Differenz des Mittelwertes der Items des computerisierten adaptiven Tests inklusive der Positionseffekte und des Mittelwertes der Items des papierbasierten Tests $Mean\left(b_{CAT} + \gamma_{k_{CAT}}\right) - Mean(b_{FIT})$.

$$b_{FIT_{M/M}} = \left(Mean\left(b_{CAT} + \gamma_{k_{CAT}}\right) - Mean(b_{FIT})\right) + b_{FIT} \qquad (26)$$

Schritt 4: Prüfung der Linkitems auf Itemparametervariation zwischen den Testformen

Die im ersten Schritt gewählten vorläufigen Linkitems werden nach der linearen Transformation aus Schritt 3 nun auf Itemparametervariation geprüft. Von den ursprünglichen Linkitems werden nach Schritt 3 nur die Items als gemeinsame Linkitems beibehalten, welche keine bzw. eine sehr geringe Variation in den Itemparametern aufweisen. Dazu wurde eine zweiseitige Prüfung auf einem Alpha-Niveau von 5 % durchgeführt. Die Nullhypothese lautet dabei: Es liegt keine Variation zwischen $b_{FIT_{M/M}}$ und $b_{CAT} + \gamma_{k_{CAT}}$ vor. Konkret wurde im Schritt 4 für jedes Item geprüft, ob sich der Wert $b_{CAT} + \gamma_{k_{CAT}}$ im Konfidenzintervall um $b_{FIT_{M/M}}$ befindet. Das Konfidenzintervall wurde um den Schwierigkeitsparameter der papierbasierten Testung gelegt, da der Wert aus dem adaptiven Test als fixer Wert angenommen wurde, der die Berichtsmetrik bildet. Sollte diese Hypothese für ein gewähltes Item abgelehnt werden, wurde das Item als Linkitem entfernt und mit den restlichen Items Schritt 3 wiederholt. Schritt 3 und Schritt 4 wurden solange wiederholt, bis für kein Linkitem die Nullhypothese mehr abgelehnt werden konnte. Es wird dabei angenommen, dass es sich bei jedem Durchlauf von Schritt 4 um eine neue Nullhypothese handelt und somit keine Alphafehler-Kumulierung vorliegt. Denn nach jedem Durchlauf wird die Nullhypothese neu aufgestellt und bezieht sich auf eine neue Stichprobe von Items.

Schritt 5: Skalierung FIT mit Ankeritems

Nach der Festlegung der Ankeritems bzw. der Linkitems wurden die Antworten des papierbasierten Tests erneut mit einem Rasch-Modell mit Ankeritems skaliert. Als Schwierigkeitsparameter der Ankeritems wird die Schwierigkeit der Items aus dem adaptiven Test inklusive der Positionseffekte ($b_{CAT} + \gamma_{k_{CAT}}$) verwendet. Die restlichen Items werden frei geschätzt. Diesmal wurde die Personenverteilung nicht auf den Mittelwert von 0 fixiert, damit die Daten der Personenverteilung der Verteilung der Berichtsmetrik (CAT) entsprechen. Der Syntax-Abschnitt dazu lautet:

```
import anchor_parameters << Anchorparameter.dat;

set constraints=none;

model Item;
```

In der Datei Anchorparameter.dat werden die Schwierigkeiten zu den entsprechenden Ankeritems hinterlegt. Der Constraint dieses Modells erfolgt durch die Festsetzung der Ankeritems. Diese Itemschwierigkeiten werden somit für die jeweiligen Items in der papierbasierten Testung angenommen, um auf derselben Metrik wie beim computerisierten adaptiven Testen berichten zu können. Als Möglichkeit, die Itempools eines computerisierten adaptiven Tests und eines papierbasierten FIT angemessen miteinander zu verbinden, bietet sich die oben vorgestellte Methode an. Die Anzahl der endgültigen Linkitems mit und ohne Korrektur der Itemparameter um die Positionseffekte kann der Tabelle 17 entnommen werden.

Tabelle 17

Anzahl der Linkitems für die Domänen Lesen (READ), Mathematik (MATH) und Naturwissenschaft (SCIE) mit und ohne Korrektur der Itemparameter um die Positionseffekte (Pos.)

Domäne	Anzahl der Linkitems			
	nach Schritt 1	nach Schritt 4 (ohne Pos.)	nach Schritt 4 (mit Pos.)	nach Schritt 4 Differenzmenge
READ	31	15	14	3
MATH	36	22	21	3
SCIE	36	28	30	6

Anmerkungen: Die Anzahl der möglichen Linkitems (nach Schritt 1) ergibt sich aus der Anzahl der Items, welche sowohl im papierbasierten FIT als auch im computerisierten adaptiven Test eingesetzt wurden. Bei der Verwendung der Linkingprozedur mit und ohne Betrachtung von Positionseffekten wird endgültig eine unterschiedliche Menge von Linkitems (Differenzmenge) ausgewählt.

Insgesamt konnten aus den gemeinsamen Items bei Berücksichtigung der Positionseffekte für Lesen 14 von 31, für Mathematik 21 von 36 und für Naturwissenschaft 30 von 36 Linkitems gewählt werden. Ohne Betrachtung der Positionseffekte werden in den Domänen Lesen und Mathematik jeweils ein Item mehr gewählt. Dieses Ergebnis ist diskutabel, da durch die Berücksichtigung der Positionseffekte eine bessere Passung der

Itemparameter zwischen den beiden Testformen erwartet wurde. Dies hätte, wie bei der Domäne Naturwissenschaft, zu einer höheren Anzahl an Linkitems bei Berücksichtigung der Itempositionseffekte führen sollen. Dennoch hat die Berücksichtigung der Positionseffekte eine bedeutsame Auswirkung auf die Wahl der Linkitems. Es werden je nach Linkingprozedur (mit oder ohne Berücksichtigung von Positionseffekten) zum Teil unterschiedliche Items als Linkitem gewählt. In der Domäne Lesen werden drei, in der Domäne Mathematik drei und in der Domäne Naturwissenschaft sechs abweichende Items (Differenzmenge) ausgewählt. D. h. beispielsweise für die Domäne Naturwissenschaft, dass sechs Items nicht in beiden Mengen an Linkitems (mit und ohne Berücksichtigung von Positionseffekten) vorhanden sind. Eine inhaltlich Erklärung dafür ist, dass ein gewähltes Linkitem nach dem Mean/Mean-Equating den gleichen Schwierigkeitsparameter beim FIT wie beim CAT aufweist. Doch nachdem die Position des Items im papierbasierten Testheft berücksichtigt wurde, stellt sich heraus, dass es einen signifikanten Unterschied im Schwierigkeitsparameter dieses Items zwischen den Testformen gibt. Andersherum kann es ebenso möglich sein, dass ein Item aus dem FIT nicht als Linkitem berücksichtig wird, da es im Schwierigkeitsparameter zu stark von der Metrik des CAT abweicht. Nachdem man aber die Position berücksichtigt, auf der das Item im papierbasierten Testheft vorgelegt wird, ergibt sich eine Gleichheit der Itemschwierigkeiten, womit das Item doch als Linkitem gewählt werden kann. Die Auswirkungen der Linkingprozedur auf die Personenparameterverteilung und die Reliabilität sind in der Tabelle 18 abgetragen. Dort sind wichtige Parameter der Personenparameterverteilung für (a) die freie Skalierung nach Schritt 1, (b) die Skalierung nach Schritt 4 ohne Betrachtung der Positionseffekte und (c) die Skalierung nach Schritt 4 mit Betrachtung der Positionseffekte zu sehen. Dabei ist zu erwähnen, dass zwischen der Skalierung mit Positionseffekten und der Skalierung ohne Positionseffekte Unterschiede in einigen Parametern zu sehen sind, welche jedoch nicht signifikant werden. Der Mittelwert der Personenverteilung verschiebt sich nach dem Linking mit Positionseffekten beispielsweise in der Domäne Lesen auf 0.334 bei einer Varianz von 0.529. Die Reliabilität ist mit .715 für einen Test mit fixer Itemreihenfolge nach 33 deutlich geringer als beim computerisierten adaptiven Testen mit .8655 (vgl. Kapitel 4.5.3). In der Domäne Mathematik verschiebt sich der Mittelwert lediglich um 0.039 Logits und in der Domäne Naturwissenschaft liegt die Kompetenz im Mittel bei 0.184. Die Reliabilitäten sind für die

Testlängen in den Domänen Mathematik (36 Items) und Naturwissenschaft (41 Items) mit über 0.8 akzeptabel.

Tabelle 18

Personenparameterverteilung der Probanden des FIT: Mittelwert (θ_{mean}), Varianzen σ_θ^2 und EAP/PV Reliabilitäten (Rel.) der Skalen für die Domänen Lesen (READ), Mathematik (MATH) und Naturwissenschaft (SCIE) für die Skalierung mit unterschiedlichen Modellen

Domäne	Skalierung	θ_{mean}	*SE*	σ_θ^2	*SE*	Rel.
READ	frei	0*	-	0.556	0.087	.735
	ohne Pos.	0.336	0.074	0.542	0.084	.713
	mit Pos.	0.334	0.075	0.529	0.083	.715
MATH	frei	0*	-	0.669	0.093	.815
	ohne Pos.	0.081	0.071	0.644	0.089	.817
	mit Pos.	0.039	0.072	0.656	0.090	.813
SCIE	frei	0*	-	0.632	0.086	.824
	ohne Pos.	0.117	0.068	0.620	0.084	.819
	mit Pos.	0.184	0.068	0.632	0.085	.823

Anmerkung: * bei der freien Skalierung wurde der Mittelwert nicht geschätzt, sondern auf $\theta_{\text{mean}} = 0$ fixiert. Aus diesem Grund gibt es dort keinen Standardfehler.

4.6.4 Zusammenfassung

Die vorgestellte Prozedur kann als ein Linking mit separater Kalibrierung der Items für beide Testformen betrachtet werden. Bei einer freien Skalierung des FIT wäre keine Verbindung zur Berichtsmetrik des adaptiven Tests möglich gewesen. Deshalb erfolgte

nach einer freien Skalierung ein Linking über die Ankeritems und die Methode des Mean/Mean-Equatings. So konnte mittels Signifikanzprüfung getestet werden, ob bei den Ankeritems in beiden Testformen auch gleiche Itemparameter vorliegen. Für Items, die sich in ihren Itemparametern zwischen den Testformen nicht signifikant unterscheiden, wurde angenommen, dass sie in beiden Testungen gleich funktionieren. Deshalb wurden diese Items als Ankeritems bei der anschließenden Skalierung des FIT auf die CAT-Parameter mit Positionseffekt fixiert. Die restlichen Items wurden frei geschätzt. Mit Hilfe der vorgestellten Prozedur können Itempositionseffekte beim Linking zweier Testformen (computerbasierter adaptiver Test und papierbasierter Test mit fester Itemreihenfolge) einfach berücksichtigt und somit einer möglichen Invarianz der Itemparameter bei unterschiedlichen Testformen vorgebeugt werden. Obwohl nur relativ kleine Positionseffekte vorliegen, sind bereits Effekte bei der Itemauswahl zu beobachten. Je nach Domäne unterscheiden sich drei bis sechs Linkitems bei Berücksichtigung der Positionseffekte im Vergleich zur Modellierung ohne Positionseffekte. Der Mittelwert der Personenverteilung verschiebt sich nach dem Linking. Die Reliabilität wird in allen drei Domänen nach dem Linking mit Positionseffekten minimal schlechter im Vergleich zur freien Skalierung ohne Linking und ohne Berücksichtigung der Positionseffekte. Bei größeren Positionseffekten sind weitreichendere Auswirkungen auf die Itemauswahl und die Personenparameterverteilung zu erwarten.

Die Methode ist für großangelegte Studien mit komplexen Testheftdesigns und bei der Verwendung computerisierter adaptiver Tests gut geeignet. In der Praxis kann beispielsweise in schwierigen Testfeldern, wo CAT nicht anwendbar ist, ein papierbasiertes FIT als Alternative angewandt und zugleich auf derselben Metrik berichtet werden. Im Unterschied zum computerisierten adaptiven Test besteht beim papierbasierten Testen die Möglichkeit, eine Aufgabe zu überspringen, was fehlende Antworten verursachen kann. Im papierbasierten Test wurden deshalb fehlende Antworten innerhalb eines Antwortblocks als falsch gewertet, wenn anschließend noch weitere Antworten auf Items gegeben wurden. Es wird davon ausgegangen, dass die Aufgaben nacheinander durchgeblättert wurden und jedes Item zumindest angeschaut wurde. In der Instruktion wurde der Hinweis gegeben, jedes Item zu beantworten und ggf. zu raten. Ein nichtbeantwortetes Item innerhalb eines Antwortblockes gilt deshalb als falsch. Fehlende

Antworten am Ende der Testung werden als fehlender Wert deklariert, da davon ausgegangen wird, dass diese Items noch nicht gesichtet wurden.

5. Zusammenfassung und allgemeine Diskussion

In dieser Arbeit wurde eine praktische Anleitung zur Konstruktion computerisierter adaptiver Tests am Beispiel der Messung schulisch erworbener Kompetenzen vorgestellt. Diese Anleitung enthält sechs Schritte. Nach einer ausführlichen Testplanung (Schritt 1) ist die Erstellung des initialen Itempools (Schritt 2) erforderlich. Dabei ist eine Kalibrierungsstudie vor der eigentlichen Testung notwendig, um Itemparameter für die untersuchte Population zu schätzen (Schritt 3). Die Anzahl der Items im Itempool sollte für die Nutzung in einem adaptiven Test möglichst groß und über die Schwierigkeitsbereiche ausgeglichen sein. Nach der Festlegung des adaptiven Algorithmus (Schritt 4) kann in einer Pilotierungsstudie das Zusammenspiel des Algorithmus und des Itempools getestet werden (Schritt 5). Da nicht in allen Anwendungsbereichen computerisierte adaptive Tests durchgeführt werden können, wird mit dem Schritt 6 eine Linkingprozedur vorgestellt. Mit dieser Prozedur können unter Berücksichtigung von Positionseffekten die Skalen aus papierbasierten Tests mit den Skalen aus computerisierten adaptiven Tests verbunden werden. In diesem Kapitel wird die vorgestellte Anleitung zur Konstruktion computerisierter adaptiver Tests diskutiert sowie deren praktischer Beitrag erläutert. Anschließend wird ein kurzer Ausblick gegeben und ein Fazit gezogen.

5.1 Diskussion und praktischer Beitrag der einzelnen Schritte

Testplanung

Mit dem ersten Schritt des vorgestellten Ansatzes wurde vor allem die Interaktion zwischen Mensch und Computer besprochen. Diese Interaktion bildet eine entscheidende Schnittstelle bei der Testung ab, an der noch viel Forschungsarbeit erfolgen kann. Durch den Computer wird das adaptive Testen im hier verwendeten Sinn erst möglich. Doch durch den Computer müssen aber im Vergleich zur klassischen papierbasierten Testung bei der Testplanung auch gänzlich neue Aspekte bedacht werden (Hartig & Klieme, 2007). Zwar ist es möglich, bei der Testplanung Simulationsstudien zu nutzen, um viele Aspekte zu prognostizieren, doch es muss auch bedacht werden, dass die simulierten Ergebnisse in der Empirie nicht immer erzielt werden können, da nie alle

Umgebungsvariablen mitmodelliert werden können (van der Linden & Glas, 2010). Der praktische Beitrag von Simulationsstudien ergibt sich vor allem daraus, dass sehr kosten- und zeitsparend bei jedem Schritt der Testentwicklung die Auswirkungen der einzelnen Entscheidungen bei der Testplanung geprüft werden können (vgl. Kapitel 3.2.2).

Da ein Schwerpunkt dieser Arbeit darauf liegt, kosten- und zeitsparend computerisierte adaptive Tests zu erstellen, ist bei der Testplanung vor allem die Nutzung vorhandener inhaltlicher Zielkonstrukte und freier Software zur Erstellung und Administration adaptiver Tests vorgeschlagen worden. Am Projekt MaK-adapt konnte beispielhaft gezeigt werden, wie eine praktische Umsetzung möglich ist. Für die Domänen Mathematik und Naturwissenschaft stellte es sich relativ einfach dar, sich auf vorhandene etablierte inhaltliche Zielkonstrukte aus anderen Studien zu stützen. Bei der Entwicklung von Tests für spezielle Bereiche (wie z. B. bei MaK-adapt für das funktionale Lesen) kann es vorkommen, dass auf kein vorhandenes Konstrukt zurückgegriffen werden kann. Zudem können komplexe inhaltliche Zielkonstrukte zwar z. B. durch Content-Balancing-Methoden in einen adaptiven Test abgebildet werden, aber bei vielen zu beachtenden Subdomänen wird der optimale Itemauswahlprozess aufgrund der Content-Balancing-Methode möglicherweise gestört. Zudem müssen für alle Inhaltsbereiche genügend Items der entsprechenden Schwierigkeiten vorliegen, um die Effizienz des adaptiven Algorithmus optimal zu unterstützen. Diese Aspekte schränken das vorgestellte Vorgehen ein und sind individuell zu handhaben. Zur Nutzung vorhandener freier Software ist anzumerken, dass diese auf lange Sicht nur dann Ressourcen spart, wenn ein offener anpassbarer Quellcode vorliegt oder wenn die Software einen guten Support zur Anpassung für die eigenen Zwecke bietet. Wenn in der Testplanung Details beschlossen werden, welche sich in der vorhandenen Software nicht abbilden bzw. mit dieser nicht durchführen lassen, ist die Programmierung einer eigenen Software zu Beginn des Projektes ratsam. Auf diese Weise besteht bestenfalls die Unabhängigkeit von Dritten und somit eine hohe Flexibilität. Die Nutzung der Software MATE hat sich für das konkrete empirische Beispiel MaK-adapt angeboten, da dort die gewünschten Itemformate verwendet werden konnten, die Software für Forschungszwecke frei zur Verfügung stand und ein enger Austausch mit dem Entwickler der Software bestand. Die in dem Kapitel Testplanung angesprochenen Aspekte zur technischen Umsetzung können

natürlich nur einen Einblick geben und nicht alle technischen Möglichkeiten und Herausforderungen abdecken. Die im Projekt MaK-adapt genutzte netzwerkbasierte Lösung mit der Speicherung des Itempools auf den lokalen Rechnern und der automatischen Übertragung der Ergebnisse über das Internet auf einen Server hatte verschiedenste Vor- und Nachteile. Allgemein wird deshalb empfohlen, stets mehrere Auslieferungsmodi parallel zu nutzen und für Notfälle eine papierbasierte Testversion mit zu entwickeln. Außerdem ist es ratsam, sich an aktuellen technischen Entwicklungen zu orientieren, um weitere Möglichkeiten der Testadministration zu nutzen (z. B. Testung über das Smartphone). Die vorgestellten Schritte im Kapitel Testplanung sind aus Autorensicht die elementarsten Bestandteile bei der Vorbereitung der Testentwicklung eines computerisierten adaptiven Tests, bevor der zweite Schritt, die Entwicklung des initialen Itempools, angegangen wird.

Entwicklung des initialen Itempools

Ein angemessener Itempool bildet die Voraussetzung dazu, im Randbereich der Kompetenzverteilung effizient messen zu können. Die Qualität des Itempools bestimmt über die Qualität des Tests (Flaugher, 2000). Bei Items mit simplen Stimulus und einfachem Antwortformat lassen sich Templates nutzen, welche automatisiert einzelne Inhalte in den Aufgaben austauschen und so kostengünstig eine große Anzahl an Items während der Testung schaffen können (Embretson, 1999). Durch die vorgestellte Methode des Itemrecycling konnten im Projekt MaK-adapt in kürzester Zeit über 300 Items entwickelt werden. Die Ergebnisse in Bezug auf die notwendige Itemanzahl und die Verteilung der Itemschwierigkeiten ist erwartungskonform. Die Items sind über die Subdomänen der inhaltlichen Zielkonstrukte und der unterschiedlichen Schwierigkeitsbereiche annähernd gleichverteilt. Es wurden in den Domänen Mathematik und Naturwissenschaft etwa 30 % mehr Items generiert, als für die Erfüllung der theoretischen Annahme notwendig waren. Im Bereich Lesen konnte aufgrund der Neuentwicklung mehrerer Items dieses Ziel nicht erreicht werden. Das Itemrecycling bietet die Möglichkeit Kosten und Zeit bei der Testentwicklung zu sparen. Ein weiterer Vorteil des Itemrecycling ist, dass genutzte Items bereits einen oder mehreren Pretest unterzogen wurden und so mit einer gesteigerten Qualität der Items, im Vergleich zur Neuentwicklung, zu rechnen ist. Der große Nachteil des Itemrecycling ist es, gerade für gute oder aufwendig entwickelte innovative Items, die Rechte zur Nutzung zu erhalten. Dieser Aspekt ist dem vorgeschla-

genen Vorgehen kritisch anzumerken. Gerade bei Studien, wo wenig Items aus anderen Studien vorhanden sind oder in denen sehr viele Items benötigt werden, wird man nicht umhin kommen, neue Items zu erzeugen. Dabei ist auch immer abzuwägen, inwiefern innovative Items (Parshall et al., 2010) eingesetzt werden müssen. Diese sind in der Konstruktion und Kalibrierung meist zeit- und kostenintensiver (Osterlind, 1998).

Pretest und Kalibrierung des Itempools

Unabhängig davon, ob neu entwickelte Items oder Items aus anderen Studien eingesetzt werden, ist zu prüfen, ob diese in der eigenen Studie angemessen funktionieren und die Items zum gewählten Modell passen. Vor allem die Prüfung der Itemqualität (Itemselektion) und die Prüfung des Modellfits sind dabei hervorzuheben (Rost, 2004). Im hier verwendeten empirischen Beispiel wurden die Items auf eine neue Gruppe von Probanden (SuS beruflicher Schulen) angewandt. Bei der Prüfung der Itemqualität wurde deshalb bei der DIF-Analyse die Zugehörigkeit zu unterschiedlichen Berufsgruppen mit berücksichtigt. Kritisch anzumerken ist, dass in dieser Arbeit auf die Berufsgruppe nur als Haupteffekt eingegangen wurde, um den Geschlechtereffekt zu untersuchen und nicht den DIF-Effekt aufgrund der Berufsgruppe selbst. Für eine ausführliche Untersuchung der DIF-Effekte auf Grundlage der Ausbildungsberufe wird auf Spoden et al. (2015) verwiesen. Die Ergebnisse zur Itemselektion sind teilweise erwartungskonform. Es konnten 93.2 % der Leseitems, 85.0 % der Mathematikitems und der 73.3 % Naturwissenschaftsitems nach der ersten Selektion beibehalten werden. Da in der Domäne Lesen schon in die Kalibrierung mit relativ wenig Items (75 Items) gearbeitet wurde, sind 68 vorhandene Items bei einer Neuentwicklung eines Tests als gut zu bewerten. In der Domäne Mathematik konnten mit 113 Items deutlich mehr als die 100 angestrebten Items und in der Domäne Naturwissenschaft mit 94 Items etwas weniger als die 100 angestrebten Items im Pool gelassen werden. Bezüglich der Studie zu den Itempositionseffekten sind die Ergebnisse dahingehend erwartungskonform, dass in den Domänen Lesen und Naturwissenschaft die Effekte mit zunehmender Testdauer größer werden. In der Domäne Mathematik zeigt sich eine Kurve, bei welcher der Positionseffekt zu Beginn und am Ende größer ist als in der Mitte der Testung. Dieses Ergebnis ist erklärungsbedürftig und benötigt weitere Untersuchungen. Eine mögliche Hypothese ist, dass mathematische Aufgaben einen abstrakteren Charakter haben können als Aufgaben aus dem Bereich Lesen und Naturwissenschaft und der Proband deshalb eine längere

Einarbeitungszeit benötigt. Weiterhin hat sich gezeigt, dass die Positionseffekte für alle Items als identisch angesehen werden können. Zu diesem Ergebnis ist anzumerken, dass in dieser Arbeit lediglich zwei unterschiedliche Multifacetten-Rasch-Modelle verglichen wurden. Dazwischen sind weitere abgestufte Modelle denkbar (Frey et al., im Druck). Für die praktische Implementation in anderen Studien kann das vorgeschlagene Vorgehen jedoch als Vorlage dienen. Es hat sich gezeigt, dass die Berücksichtigung von Positionseffekten auf die Personenverteilung der Kalibrierungsstudie keine direkte Auswirkung hat. D. h., an der Varianz und Reliabilität sind die Effekte nicht zu erkennen. Aber gerade auf den wichtigen Schritt der Itemselektion hat das Vorliegen von Positionseffekten Auswirkungen.

Die Kalibrierungsstudie ist außerdem dahingehend zu diskutieren, ob die Festlegung der Itemparameter, welcher der spätere adaptive Algorithmus benötigt, angemessen erfolgt ist. In dieser Arbeit wurde die Bedeutung des Testheftdesigns hervorgehoben. Gerade in Bezug auf adaptives Testen muss sich der Testentwickler die Frage stellen: Wie kann der Itemparameter geschätzt werden, wenn er durch ein festes Testheft ermittelt wurde, aber später adaptiv vorgelegt wird? Es wurde deshalb der Zusammenhang zwischen dem Testheftdesign und der Ermittlung möglicher Positionseffekte hergestellt. Das vorgestellte Testheftdesign bietet zwar Vorteile, Positionseffekte auf Einzelpositionen zu ermitteln und Rückmeldungen von Probanden zu allen drei Domänen zu erhalten. Dennoch wäre, gerade in Anbetracht der geringen Anzahl an Antworten auf ein Item pro Position auch ein weniger komplexes Testheft denkbar gewesen, in welchem von Beginn an Positionsstufen gebildet werden. Zudem ist für die Entwicklung von unidimensionalen Tests auch ein Design möglich, in welchem jeder Proband nur eine Domäne bei der Kalibrierung erhält. Damit erspart sich der Testentwickler mögliche Folgeeffekte der vorhergehenden Domäne bei der Analyse der Positionseffekte und der Ermittlung der Itemparameter.

CAT – Algorithmus

Nachdem ein kalibrierter Itempool vorliegt, können im vierten Schritt die unterschiedlichen Elemente des adaptiven Algorithmus festgelegt werden. Die vorgestellten Elemente des adaptiven Algorithmus (Startpunkt, Itemauswahl, Fähigkeitsschätzung, Testende und Restriktionen) sind häufig die Aspekte, die in der Literatur unter dem

Stichwort CAT behandelt werden (van der Linden & Glas, 2010). Darüber hinaus gibt es aus praktischer Sicht jedoch zu jedem Element verschiedene Wahlmöglichkeiten, welche abhängig von der verwendeten Software und den entsprechenden Zielen der Studie sind (z. B. kurze Testungen, geringer Messfehler, hohe Testsicherheit, exakte Abbildung des theoretischen Zielkonstrukts usw.). Das Projekt MaK-adapt hatte vorrangig die Ziele, einen möglichst kurzen Test zu erstellen, welcher auch in den Randbereichen der Kompetenzverteilung gut differenziert und die Subdomänen des inhaltlichen Zielkonstrukts gleichmäßig abbildet. Um einen kurzen Test mit geringer Testzeit zu erhalten, wurde für die Pilotierungsstudie die Itemanzahl auf max. 48 Items und die Testzeit auf max. 40 Minuten festgelegt. Kritisch dabei ist, dass diese Einstellungen bei der Pilotierungsstudie für alle drei Domänen identisch waren. Die Bearbeitungszeit der einzelnen Items unterschied sich jedoch zwischen den Domänen. Dieser Aspekt wurde nach der Pilotierungsstudie dadurch berücksichtigt, dass je nach Domäne in gleicher Testzeit unterschiedlich viel Items beantwortet werden konnten. Die Überprüfung der Einstellung für den adaptiven Algorithmus fand vor der Pilotierungsstudie durch eine erste Simulationsstudie statt. Dabei wurde u. a. die Reliabilität in Bezug zur Testlänge und die gleichmäßige Verteilung der Items über die Subdimensionen durch den MPI geprüft. Die Ergebnisse bezüglich gleichmäßiger Verteilung der Items und erwarteter Reliabilität waren erwartungskonform. Das Ziel, im Randbereich der Kompetenzverteilung angemessen zu messen, wurde weitestgehend erreicht. Der *SE* ist im Randbereich der Kompetenzverteilung zwar leicht höher als in der Mitte der Kompetenzverteilung. Dennoch sind die *SE* des computerisierten adaptiven Tests im Randbereich deutlich geringer als beim FIT. Dieses Ergebnis entspricht den Erwartungen (Weiss, 2016). Beim Hinzufügen neuer Items zum Itempool sollten vor allem Items für den Randbereich (schwierige bzw. leichte Items) produziert werden, um den adaptiven Algorithmus beim Itemauswahlprozess optimal zu unterstützen. Kritisch kann für das Projekt MaK-adapt angemerkt werden, dass es keine Kontrolle der Häufigkeit vorgelegter Items gibt (Exposure-Control). Da die Testsicherheit im Projekt MaK-adapt aber keine wesentliche Rolle spielte, wurde dies bewusst außer Acht gelassen. Die Wahl des BME als Personenparameterschätzer, die Annahme des Startwertes für θ von 0 und die zufällige Wahl des Startitems aus 10 passenden Items mit mittlerer Schwierigkeit kann ebenfalls diskutiert werden. Der BME bot sich als Schätzer an, da so das Vorwissen über die Verteilung aus der Kalibrierungsstudie bei der Schätzung genutzt werden konnte. Gerade bei sehr

kurzen Testungen hat dies einen Vorteil. Sollten die Tests aber später in einer abweichenden Kohorte Anwendung finden und individuelle Rückmeldungen angestrebt werden, sollte ggf. ein anderer Schätzer genutzt werden (Wainer & Mislevy, 2000). Bei Rückmeldung auf individueller Ebene ist zudem zu bedenken, eventuell das Testende so einzustellen, dass erst ab einem Erreichen eines bestimmten SE der Test abgebrochen wird. Somit sind die Ergebnisse zwischen den Testpersonen besser vergleichbar. Bei Studien mit Vorwissen über die konkreten Probanden sollte der Startwert von θ entsprechend angepasst werden. Da kein Vorwissen über die Fähigkeit des konkreten Probanden zu Testbeginn vorlag, wurde ein Startwert für θ von 0 angenommen. Um nicht bei jedem Test mit dem gleichen Item zu beginnen, wurde das Startitem zufällig aus den 10 besten Items (maximale Information) für das gegebene θ von 0 gezogen. So konnte dem Bekanntwerden der Iteminhalte etwas vorgebeugt werden. Zudem taucht dadurch in einem Computerpool bei einer Testung mehrerer Personen nicht auf jedem Monitor zu Beginn dasselbe Item auf.

Insgesamt kann der vierte Schritt dieser Anleitung kritisch betrachtet werden. Aus praktischer Sicht sind die Einstellungen am adaptiven Algorithmus bei Bedarf neu festzulegen und nicht endgültig festgelegt. Diese Kritik betrifft insgesamt jedoch den gesamten Ablauf der vorgestellten Anleitung und somit den im Kapitel 4.5.4 theoretisch vorgestellten Ablauf zur Wartung und Pflege eines computerisierten adaptiven Tests. Bei einer Änderung eines Tests, z. B. aufgrund einer Wartungsarbeit, sollte in der Regel der vorgestellte Ablauf ab dem Schritt der Änderung erneut durchlaufen werden. Konkret heißt das beispielsweise, dass bei Änderungen des theoretischen Zielkonstrukts ab dem ersten Schritt , bei Änderungen am Itempool ab dem zweiten Schritt, bei Bekanntwerden von Itemparameterdrift ab dem dritten Schritt oder bei Änderungen am Algorithmus ab dem vierten Schritt dieser und alle weiteren Schritte zu durchlaufen sind. Dabei ist bei jedem Schritt zu prüfen, ob eine erneute Durchführung notwendig ist. Bei einer Elimination eines Items muss nicht gleich eine neue Kalibrierung mit allen Folgeschritten erfolgen. Alle Änderungen am computerisierten adaptiven Test haben jedoch zur Folge, dass der Test in seiner neuen Form veröffentlicht und angewendet wird.

CAT – Veröffentlichung und Anwendung

Im Projekt MaK-adapt wurde in der Pilotierungsstudie das Zusammenspiel des Itempools mit dem adaptiven Algorithmus in der Software MATE und der netzwerkbasierten Lösung erstmals empirisch im Feld erprobt. Änderungen nach der Pilotierungsstudie, z. B. am Itempool oder dem Algorithmus, führen anschließend stets wieder zur Durchführung des fünften Schritts. Zudem ergeben sich nach der Veröffentlichung und der Anwendung der computerisierten adaptiven Tests im Feld häufig neue Erkenntnisse bezüglich des Algorithmus und des Itempools, die anschließend wieder einfließen. Aus diesem Grund wurde die Methode zur Wartung und Pflege diesem Schritt zugeordnet. Die Pflege der Itempools, der Software und der Skalen beinhaltet sowohl die Themen Testsicherheit und Prüfung von Itemparameterdrift als auch das kontinuierliche Hinzufügen und Entfernen von Items. Es ist deshalb naheliegend, dass nach Abschluss des Projektes MaK-adapt noch einmal minimale Änderungen am Itempool erfolgten. An den genannten Änderungen ist kritisch anzumerken, dass bisher keine weiteren Items dem Itempool hinzugefügt wurden. Der Itempool für die Domäne Mathematik enthält abschließend 102 Items, der Itempool für die Domäne Lesen 62 Items und der Itempool für die Domäne Naturwissenschaft 94 Items. Diese Zahlen sind nach den angestrebten Zahlen von 25 Items pro Subdomäne (fünf Items pro Schwierigkeitsbereich pro Subdomäne), wie bereits erwähnt, lediglich für die Domäne Mathematik erwartungskonform. Eine konkrete Änderung im Projekt MaK-adapt nach der Pilotierungsstudie ergab sich bezüglich der Abbruchkriterien. Dazu wurden die simulierten Ergebnisse aus einer zweiten Simulationsstudie in der Software MATE zu den Abbruchkriterien in Zusammenhang mit der empirischen Reliabilität aus der Pilotierungsstudie verglichen. Es ist hervorzuheben, dass die Reliabilität aus den Simulationen für die Domänen Lesen und Mathematik beinahe identisch mit der Reliabilität auf Grundlage der empirischen Ergebnisse aus der Pilotierungsstudie sind. Für die Domäne Naturwissenschaft liegt die empirische Reliabilität leicht unter der simulierten. Die Ergebnisse der Pilotierungsstudie entsprechen somit den Erwartungen und zeigen, dass auf Basis der Itempools eine hinreichende Messpräzision mit geringer Itemanzahl beim adaptiven Testen erzielt werden kann. Zudem kann der MPI als zuverlässiges Instrument des Content-Balancing beim computerisierten adaptiven Testen empfohlen werden, wenn der Itempool relativ gleichverteilt angelegt wird. Die Ergebnisse der veröffentlichten Tests aus der Pilotie-

rungsstudie werden auf der Logitskala abgebildet (vgl. Tabelle 11 auf S. 141). Hier wäre zukünftig eine Standardisierung der Verteilungen aus den drei Domänen auf einheitliche Mittelwerte und Standardabweichungen sowie eine qualitative Interpretation der quantitativen Ergebnisse, z. B. durch sogenannte Kompetenzstufen, ähnlich wie bei PISA möglich (Heine, Sälzer, Borchert, Sibberns & Mang, 2013).

Die Prüfung der verwendeten Software MATE und der netzwerkbasierten Lösung im Zusammenhang mit der genutzten Hardware an den beruflichen Schulen zeigte Nachteile des gewählten Vorgehens. Die Tests konnten aufgrund von technischen Problemen und fehlender Computerausstattung nicht an allen gewünschten Schulen eingesetzt werden. Eine Offline-Version wurde leider erst nach der Pilotierungsstudie entwickelt. Durch einen parallelen Einsatz beider Versionen hätten deutlich mehr Personen getestet werden können. Der Einsatz einer papierbasierten Version als Alternative hat sich deshalb als sinnvoll herausgestellt. Nach der Projektlaufzeit steht lediglich noch die Offline-Version (Bernhardt et al., 2016) der drei computerisierten adaptiven Tests und das papierbasierte Testheft zur Verfügung. Die vorgestellte papierbasierte Version ist dahingehend kritisch zu beurteilen, dass jede Domäne genau ein Testheft mit einer Itemzusammenstellung enthält. Hier wäre bezüglich Testsicherheit eine größere Palette an Testheften wünschenswert. Die Online-Version wurde aufgrund der fehlenden Mittel nach Projektende nicht mehr weiter unterstützt. Insgesamt hat sich durch die Auswertung der Pilotierungsstudie auch gezeigt, dass die Einstellungen des adaptiven Algorithmus aus Schritt 4, bis auf das Abbruchkriterium, beibehalten werden können.

Linking mit papierbasierter Testung

Die Nutzung der vorgestellten papierbasierten Testversion als Alternative ist dann sinnvoll, wenn die Ergebnisse auf die vorhandene Skala aus dem computerisierten adaptiven Test gelinkt werden können. Dazu wurde eine mögliche Methode vorgestellt, bei der Itempositionseffekte berücksichtigt werden. Bisher wird das Linking nicht als Schritt zur Erstellung eines computerisierten adaptiven Tests gesehen. In dieser Arbeit wird das Linking als letzter Schritt eingeführt, um standardmäßig eine Schnittstelle zum FIT zu erhalten. Dabei soll noch einmal betont werden, dass die Herstellung einer Äquivalenz zweier Skalen bei der Verwendung von unterschiedlichen Testmedien nur schwer herzustellen ist. Aus diesem Grund wurde hier als Linking eine Kalibrierung mit

Ankeritems vorgeschlagen, so dass die Skala des papierbasierten Tests und die Skala des computerisierten adaptiven Tests miteinander verbunden werden konnten. Die vorgestellte Prozedur kann als ein Linking mit separater Kalibrierung der Items für beide Testformen betrachtet werden. Mit Hilfe der vorgestellten Prozedur können Itempositionseffekte beim Linking zweier Testformen (computerbasierter adaptiver Test und papierbasierter Test mit fester Itemreihenfolge) einfach berücksichtigt und somit einer möglichen Invarianz der Itemparameter bei unterschiedlichen Testformen vorgebeugt werden.

Nicht erwartungskonform war der relativ geringe Anteil an gefundenen Linkitems. Zwar hat die Nutzung von Itempositionseffekten gezeigt, dass teilweise andere Items als Linkitems verwendet werden als mit Positionseffekten. Doch es wurde erwartet, dass nach Betrachtung dieser Effekte deutlich mehr Linkitems vorhanden sind, als ohne Betrachtung dieser Effekte. Dieser Zuwachs wurde nicht beobachtet, was vermutlich aufgrund der geringen Positionseffekte zurückzuführen ist. In der Domäne Lesen hat sich zudem gezeigt, dass weniger als die Hälfte der Items nach der Linkingprozedur als Ankeritems ausgewählt worden sind. Dies kann darauf deuten, dass gerade bei langen Items mit viel Text, die Items bei papierbasierten Testungen mit fester Itemreihenfolge anders funktionieren als bei adaptiven Testungen am Computer. Was dagegen spricht ist, dass die gewählten Linkitems in ihren Eigenschaften (Länge Text, Anzahl Bilder, Tabellen usw.) nicht anders sind als die nicht gewählten Items. Insgesamt kann die vorgestellte Methode als Vorlage für ein Linking unter den genannten Voraussetzungen empfohlen werden.

5.2 Ausblick

Die vorgestellte praktische Anleitung bietet eine Möglichkeit zur Erstellung von computerisierten adaptiven Tests. Sie ist als Ausgangspunkt zu diskutieren und durch weitere Analysen zukünftig zu erweitern. Im Bildungsbereich, beispielsweise bei großangelegten Vergleichsstudien, kann die Nutzung computerisierter adaptiver Tests zur Verbesserung der genutzten Methoden beitragen (Luecht, 2013). Da sich die im empirischen Beispiel dieser Arbeit verwendeten inhaltlichen Zielkonstrukte in den Domänen Mathematik und Naturwissenschaft an die theoretischen Rahmen von PISA

und TIMSS anlehnen, kann die vorgestellte Studie im weitesten Sinne als Machbarkeitsstudie für den Einsatz von computerisierten adaptiven Testen bei PISA oder TIMSS oder aber zur Überprüfung der Bildungsstandards gesehen werden (Frey & Ehmke, 2008).

Als ein möglicher nächster Schritt zur Erweiterung dieser Arbeit kann die Implementation von Itempositionseffekten im adaptiven Algorithmus empirisch angewendet und geprüft werden. Gerade in Studien bei Vorliegen von größeren Positionseffekten als im Projekt MaK-adapt stellt die Berücksichtigung dieser Effekte eine Möglichkeit zur Verbesserung der Itemauswahl und der Personenparameterschätzung dar. Vor einer empirischen Studie sollten mögliche Effekte auf die Itemauswahl und Personenparameterschätzung durch Simulationsstudien in Abhängigkeit von der Größe der Positionseffekte geprüft werden. Zudem könnte in einer weiteren umfangreicheren Arbeit auf andere Möglichkeiten des Testheftdesigns bei der Kalibrierungsstudie, auf die empirische Prüfung weiterer adaptiver Softwarepakete sowie anderer Auslieferungsmodi bei der Veröffentlichung des computerisierten adaptiven Tests eingegangen werden. Hier erscheinen vor allem Lösungswege sinnvoll, in denen über alternative Hardware der haptischen Steuerung, z. B. über Tablets mit Touch-Screen oder elektronischen Stiften, der papierbasierten Testung näher gekommen wird. In dem Zusammenhang sind weiterführende empirische Studien zur Prüfung der unterschiedlichen Auswirkungen computerisierten adaptiven Testens im Vergleich zum papierbasierten FIT wünschenswert, z. B. in Bezug zur Motivation (Asseburg, 2011), zur Testsicherheit (Frey & Ehmke, 2008) oder zu affektiven Zuständen bei der Testbearbeitung (Watson, Clark & Tellegen, 1988). Für die Entwicklung eines adaptiven Tests unter Berücksichtigung von komplexeren IRT-Modellen (2PL, 3PL), multidimensionalen Modellen oder weiteren denkbaren Testmodellen (vgl. Kapitel 3.1.1) sowie zur Nutzung polytomer Items oder nominaler Antwortmodelle (Ayala, 1992) können in zukünftigen Arbeiten praktische Schritte beschrieben und diese empirisch geprüft werden.

5.3 Fazit

Die praktische Testentwicklung eines IRT-basierten computerisierten adaptiven Tests ließ zu Beginn dieser Arbeit einige Fragen offen. Zum Thema CAT stehen zwar verschiedenste Werke zur Verfügung (Hambleton et al., 1991; Meijer & Nering, 1999;

Thompson & Weiss, 2011; van der Linden & Glas, 2010; Wainer & Dorans, 2000; Wise & Kingsbury, 2000), doch es gab kein dem Autor bekanntes zusammenfassendes Werk zur praktischen Erstellung computerisierter adaptiver Tests. Im Vergleich zu dem Verständnis der Testentwicklung eines adaptiven Tests als Baukastenprinzip mit den Bausteinen Itempool, Startpunkt, Itemauswahl, usw. (Thompson & Weiss, 2011; Wainer & Dorans, 2000; Wise & Kingsbury, 2000) ist diese Arbeit erstmals als ein umfangreicher praktischer Rahmen zur Erstellung eines computerisierten adaptiven Tests zu sehen. Anhand der empirischen Untersuchungen im Projekt MaK-adapt konnte beispielhaft gezeigt werden, wie mit wenigen Mitteln in kurzer Zeit die notwendigen Schritte zur Entwicklung eines computerbasierten adaptiven Tests in einem neuen heterogenen Feld durchgeführt werden können. Als weiterer Schritt bei der Erstellung eines computerisierten adaptiven Tests wurde ein möglicher Ansatz zur Berücksichtigung von Positionseffekten eingeführt. Die Anwendbarkeit der vorgestellten Methode wurde empirisch erprobt und geprüft. Ergänzend wurde anhand empirischer Daten aus dem Projekt MaK-adapt beispielhaft die Auswirkung von Positionseffekten bei der Kalibrierungsstudie gezeigt. Außerdem wurde ein Ansatz zur Verbindung (Linking) der Metrik eines Tests mit fester Itemreihenfolge an die Metrik eines adaptiven Tests vorgestellt. Dieser Ansatz wurde empirisch am Projekt MaK-adapt angewandt und geprüft. Itempositionseffekte können dadurch zukünftig bei der Kalibrierung der Itemparameter berücksichtigt und eine innovative Methode zum Linking von Tests angewandt werden.

Diese Arbeit trägt dazu bei, dass Testentwickler zukünftig häufiger den Schritt wagen, trotz knapper Ressourcen, einem heterogenen Untersuchungsfeld und geringem Vorwissen über die Testentwicklung von adaptiven Tests, einen computerisierten adaptiven Test zu entwickeln. Somit werden, unabhängig von den weiteren Untersuchungsmöglichkeiten, durch diese umfangreiche empirisch geprüfte Handreichung zur Erstellung von computerisierten adaptiven Tests, die Nutzungsmöglichkeiten und der Anwenderkreis computerisierten adaptiven Testens erweitert.

6. Literaturverzeichnis

Achtenhagen, F. & Baethge, M. (2008). Kompetenzdiagnostik als Large-Scale-Assessment im Bereich der beruflichen Aus- und Weiterbildung. In Prenzel, M., Gogolin, I., & Krüger, H.-H. (Hrsg.), *Kompetenzdiagnostik* (S. 51-70). Wiesbaden: VS Verlag für Sozialwissenschaften.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31* (2-3), 162-172.

Adams, R. J., Wu, M. L., Haldane, S. & Sun, X. (2012) ACER ConQuest [Computer software]. Melbourne: ACER Press.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* (30), 9-14.

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50* (4), 408-426.

Alexandrowicz, R. & Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: a simulation study and an application. *Psychology Science Quarterly, 50* (1), 64-74.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika* (38), 123-140.

Arendasy, M., Hornke, L. F., Sommer, M., Häusler, J., Wagner-Menghin, M., Gittler, G. et al. (2009). *Intelligenz-Struktur-Batterie (INSBAT): Eine Testbatterie zur Messung von Intelligenz (Manual & Test).* Mödling, Österreich: Schufried GmbH.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests*. Dissertation, Christian-Albrechts-Universität zu Kiel. Kiel.

Ayala, R. J. de. (1992). The Nominal Response Model in Computerized Adaptive Testing. *Applied Psychological Measurement, 16* (4), 327-343.

Beck, K., Landenberger, M. & Oser, F. (Hrsg.). (2016). *Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT* (Wirtschaft - Beruf - Ethik, Bd. 32). Bielefeld: WBV.

Bernhardt, R., Balkenhol, A., Ebermann, C., Frey, A., Seeber, S. & Ziegler, B. (2013). *Nutzung der adaptiven Tests zur Messung allgemeiner Kompetenzen im Rahmen der ASCOT-Initiative - Manual. (MaK-adapt-Manual)*. Jena: Friedrich-Schiller-Universität.

Bernhardt, R., Frey, A., Ziegler, B. & Seeber, S. (2016). *Nutzung der offline-Version zur Administration von drei adaptiven Tests zur Messung allgemeiner Kompetenzen in Lesen, Mathematik und Naturwissenschaft - Manual.* Jena: Friedrich-Schiller-Universität.

Birnbaum, A. (1968). Estimation of an ability. In Lord, F. M., & Novick, M. R., F. M. Lord, M. R. Novick & A. Birnbaum (Eds.), *Statistical theories of mental test scores* (Addison-Wesley series in behavioral science quantitative methods, vol. 4310, 2nd ed., pp. 423-479). Reading, MA: Addison-Wesley Educational Publishers Inc.

Boo, J. & Vispoel, W. P. (1998, April). *Computer versus paper-and-pencil assessment of educational development: score comparability and examinee preferences*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Born, S. & Frey, A. (2016). Heuristic Constraint Management Methods in Multidimensional Adaptive Testing. *Educational and Psychological Measurement*.

Bowles, R. P., Wise, S. L. & Kingsburry, G. G. (2008). *A report on position effects in the NCLEX RN examination*. Chicago: Joint Research Council of the National Council of State Boards of Nursing.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analystical extensions. *Psychometrika* (52), 345-370.

Cheng, Y. & Chang, H.-H. (2009). The maximum priority index method for severely item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62,* 396.

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33* (8), 644-645.

Clauser, B. E. & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice, 17* (1), 31-44.

Colton, G. D. (1998). Exam security and high-tech cheating. *The Bar Examiner, 67* (3), 13-35.

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects.* Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Davey, T. & Lee, Y. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE® revised general test*. Princeton, N.J.: Educational Testing Service.

Davis, J. & Ferdous, A. (2005). *Using Item Difficulty and Item Position to Measure Test Fatigue*. Washington, DC: American Institutes for Research.

Dawis, R. V. & Whitely, S. E. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36* (2), 329-337.

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50* (2), 164-185.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer & N. J. Dorans (Eds.), *Computerized adaptive testing. A primer* (2nd ed., pp. 135-158). Mahwah, N.J.: Lawrence Erlbaum Associates.

Dorans, N. J., Pommerich, M. & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Economides, A. A. & Roupas, C. (2007). Evaluation of Computer Adaptive Testing Systems. *International Journal of Web-Based Learning and Teaching Technologies, 2* (1), 70-87.

Eggen, T. J. H. M. & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica, 32,* 107-132.

Eignor, D. R. & Stocking, M. L. (1986). *An investigation of possible causes for the inadequacy of IRT pre-equating*. Princeton, N.J.: Educational Testing Service.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64* (4), 407-433.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists* (Multivariate applications book series). Mahwah, N.J: L. Erlbaum Associates.

Flaugher, R. (2000). Item Pools. In H. Wainer & N. J. Dorans (Hrsg.), *Computerized adaptive testing. A primer* (2nd ed, S. 37-59). Mahwah, N.J.: Lawrence Erlbaum Associates.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Auflage, S. 275-293). Heidelberg: Springer.

Frey, A., Bernhardt, R. & Born, S. (im Druck). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests. *Diagnostica*.

Frey, A. & Ehmke, T. (2008). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik: Zeitschrift für Erziehungswissenschaft* (S. 169-184). Wiesbaden: VS Verlag für Sozialwissenschaften.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice, 28* (3), 39-53.

Frey, A. & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35* (2-3), 89-94.

Frey, A. & Seitz, N.-N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the programme for international student assessment. *Educational and Psychological Measurement, 71* (3), 503-522.

Glas, C. A. W. (2010). Item Parameter Estimation and Item Fit Analysis. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of Adaptive Testing* (Statistics for social and behavioral sciences, pp. 269-288). New York: Springer.

Green, B. F. (2000). System Design and Operation. In H. Wainer & N. J. Dorans (Hrsg.), *Computerized adaptive testing. A primer* (2nd ed, S. 23-36). Mahwah, N.J.: Lawrence Erlbaum Associates.

Guo, F. & Wang, L. (2005). *Evaluating scale stability of a computer adaptive testing system. McLean, VA: GMAC.* (Research Report). McLean, VA: GMAC.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed). Mahwah, N.J: Lawrence Erlbaum Associates.

Hambleton, R. K. (1982). *Item response theory: The three-parameter logistic model (CSE Report 220)*. Los Angeles: Center for the Study of Evaluation, University of California.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Zaal, J. N. & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications..* New York, NY: Kluwer Academic Publishers.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31* (5), 457-459.

Harris, D. J. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15* (3), 247-256.

Hartig, J. & Buchholz, J. (2012). A multilevel item repsonse model for item position effetcs and individual persistence. *Psychological Test and Assessment Modeling, 54* (4), 418-431.

Hartig, J. & Klieme, E. (Hrsg.). (2007). *Möglichkeiten und Voraussetzungen technologie-basierter Kompetenzdiagnostik* (Bildungsforschung, Band 20). Bonn.

Heine, J.-H., Sälzer, C., Borchert, L., Sibberns, H. & Mang, J. (2013). Technische Grundlagen des fünften internationalen Vergleichs. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *Pisa 2012. Fortschritte und Herausforderungen in Deutschland* (S. 309-346). Münster: Waxmann.

Holland, P. W. (2007). A Framework and History for Score Linking. In N. J. Dorans, M. Pommerich & P. W. Holland (Hrsg.), *Linking and aligning scores and scales* (S. 5-30). New York: Springer.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning.* Hillsdale: Erlbaum.

Johanson, G. A. & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement, 70* (3), 394-400.

Kahlecke, A. (2014). *Vergleich der Leistungsmotivation zur Testbearbeitung zwischen einem computerisierten adaptiven Test und einem papierbasierten Test mit fester Itemreihenfolge im Bereich der Kompetenzmessung an berufsbildenden Schulen.* (Nicht veröffentlichte Bachelorarbeit). Jena: Friedrich-Schiller-Universität.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77* (1), 153-162.

Kingston, N. M. & Dorans, N. J. (1984). Item location effetcs and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8* (2), 147-154.

Knigge, J. (2011). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz "Musik wahrnehmen und kontextualisieren"* (Empirische Forschung zur Musikpädagogik, Bd. 2). Münster: LIT Verlag.

Kolen, M. J. (2007). Data Collection Designs and Linking Procedures. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31-55). New York: Springer.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking. Methods and practices* (Statistics for social and behavioral sciences, 3. ed). New York, NY: Springer.

Kolen, M. J. & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27* (1), 27-39.

Kröhne, U. & Frey, A. (2013). *Multidimensional Adaptive Testing Environment (MATE): Manual.* Frankfurt a. M.: German Institute for International Educational Research.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50* (3), 311-327.

Leary, L. F. & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55* (3), 387-413.

Li, F., Cohen, A. & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement, 49* (4), 362-379.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In Chae, S., Kang, U., Jeon, E. & Linacre, J. M. (Ed.), *Development of computerized middle school achievement test. MESA Memorandum No. 69.* (pp. 1-58). Seoul, South Corea: Komesa Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, N.J: Lawrence Erlbaum Associates.

Loyd, B. H. & Hoover, H. D. (1980). Vertical Equating Using the Rasch Model. *Journal of Educational Measurement, 17* (3), 179-193.

Luecht, R. M. (2013). Computer-Baed and Computer-Adaptive Testing. In M. Simon, K. Ercikan & M. Rousseau (Eds.), *Improving large-scale assessment in education. Theory, issues, and practice* (pp. 62-86). New York: Routledge.

Martin-Löf, P. (1977). Exact Tests, Confidence Regions and Estimates. *Synthese, 36* (2), 195-206.

Meijer, R. R. & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23* (3), 187-194.

Meyers, J. L., Miller, G. E. & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22* (1), 38-60.

Meyers, J. L., Murphy, S., Goodman, J. & Turhan, A. (2012, April). *The impact of item position change on item parameters and common equating results under the 3PL model*. Presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, B.C.

Miller, G. E. & Fitzpatrick, S. J. (2008). Expected Equating Error Resulting From Incorrect Handling of Item Parameter Drift Among the Common Items. *Educational and Psychological Measurement, 69,* 357-368.

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Auflage, S. 215-259). Heidelberg: Springer.

Moses, T., Yang, W. & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement, 44* (2), 157-178.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y. & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nydick, S. W. (2014). catIrt: An R Package for Simulating IRT-Based Computerized. R package Version 0.4-2. Verfügbar unter http://CRAN.R-project.org/package=catIrt

OECD. (2009). *PISA 2009 Assessment Framework. Key competencies in reading, mathematics and science*. Paris: OECD.

Orlando, M. & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement, 24* (1), 50-64.

Osterlind, S. J. (1998). *Constructing test items. Multiple-choice, constructed-response, performance, and other formats* (Evaluation in education and human services, 2nd ed). Boston: Kluwer Academic Publishers.

Osterlind, S. J. & Everson, H. T. (2009). *Differential Item Functioning*. Thousand Oaks, CA: SAGE Publications.

Parshall, C. G., Harmes, J. C., Davey, T. & Pashley, P. J. (2010). Innovative Items for Computerized Testing. In W. J. van der Linden & C. A. Glas (Hrsg.), *Elements of Adaptive Testing* (Statistics for social and behavioral sciences, S. 215-230). New York: Springer.

Pohl, S. & Carstensen, C. H. (2012). *NEPS technical report - Scaling the data of the competence tests (NEPS working paper no. 14)*. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pommerich, M. & Harris, D. J. (2003, April). *Context effects in pretesting: Impact on item statistics and examinee scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reckase, M. D. (2009). *Multidimensional Item Response Theory* (Statistics for social and behavioral sciences). New York: Springer.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52* (2), 127-141.

Rost, J. (1990). Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. *Applied Psychological Measurement, 14* (3), 271-282.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (Psychologie Lehrbuch, 2., vollst. überarb. und erw. Aufl.). Bern: Huber.

Rost, J. (2006). Item-Response-Theorie. In Petermann, F. & Eid, M. (Hrsg.), *Handbuch der Psychologischen Diagnostik* (Handbuch der Psychologie, Bd. 4, S. 261-274). Göttingen: Hogrefe.

Rudner, L. M. (2010). Implementing the Graduate Management Admission Test Computerized Adaptive Test. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of Adaptive Testing* (Statistics for social and behavioral sciences, pp. 151-165). New York: Springer.

Schwarz, M. (1978). Estimationg the dimensions of a model. *Annals of Statistics* (6), 461-464.

Schweizer, K., Troche, S. J. & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50* (8), 1249-1254.

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly, 51* (1), 47-64.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61* (2), 331-354.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Hrsg.), *Encyclopedia of social measurement* (Bd. 1, S. 429-438). New York, NY: Academic Press.

Spoden, C., Frey, A., Bernhardt, R., Seeber, S., Balkenhol, A. & Ziegler, B. (2015). Differenzielle Domänen- und Itemeffekte zwischen Ausbildungsberufen bei der Erfassung allgemeiner schulischer Kompetenzen von Berufsschülerinnen und Berufsschülern. *Zeitschrift für Berufs- und Wirtschaftspädagogk* (2), 168-188.

Steinberg, L., Thiessen, D. & Wainer, H. (2000). Validity. In H. Wainer & N. J. Dorans (Hrsg.), *Computerized adaptive testing. A primer* (2nd ed, S. 184-229). Mahwah, N.J.: Lawrence Erlbaum Associates.

Strain-Seymour, E., Walter, D. W. & Robert, P. D. (2009). *Strategies and Processes for Developing Innovative Items in Large-Scale Assessments*. (Research Report). Iowa City, IA: Pearson Education.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer & N. J. Dorans (Hrsg.), *Computerized adaptive testing. A primer* (2nd ed, S. 159-184). Mahwah, N.J.: Lawrence Erlbaum Associates.

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16* (1), 1-9.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14* (2), 181-196.

Vale, C. D. (1986). Linking Item Parameters Onto a Common Scale. *Applied Psychological Measurement, 10* (4), 333-344.

Vale, C. D. (2006). Computerized Item Banking. In S. M. Downing & T. M. Haladyna (Hrsg.), *Handbook of test development* (S. 261-286). Mahwah, N.J.: L. Erlbaum.

Van der Linden, W. J. & Glas, C. A. (Hrsg.). (2010). *Elements of Adaptive Testing* (Statistics for social and behavioral sciences). New York: Springer.

Van der Linden, W. J. & Hambleton, R. K. (2005). *Linear models for optimal test design* (Statistics for social and behavioral sciences). New York, NY: Springer.

Wainer, H. & Dorans, N. J. (Hrsg.). (2000). *Computerized adaptive testing. A primer* (2nd ed). Mahwah, N.J.: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (2000). ltem response theory, item calibration, and proficiency estimation. In H. Wainer & N. J. Dorans (Eds.), *Computerized adaptive testing. A primer* (2nd ed., pp. 61-99). Mahwah, N.J.: Lawrence Erlbaum Associates.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54* (3), 427-450.

Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54* (6), 1063-1070.

Weirich, S., Hecht, M. & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement,* 1-14.

Weiss, D. J. (2016). Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences, 6* (1).

Weiss, D. J. & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Assessment Systems Corporation.

Wise, S. L. & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica* (21), 135-155.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest version 2.0. Generalised item response modelling software*. Camberwell, Vic.: ACER Press.

Yan, D., Lewis, C. & Stocking, M. (2002). Adaptive Testing Without IRT in the Presence of Multidimensionality. *ETS Research Report Series, 2002* (1), i-27.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17* (4), 297-311.

Yoes, M. E. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model.* Saint Paul, MN: Assessment Systems Corporation.

Ziegler, B., Balkenhol, A., Keimes, C. & Rexing, V. (2012). Diagnostik „funktionaler Lesekompetenz". *bwp@ Berufs- und Wirtschaftspädagogik–online* (22), 1-19.

Ziegler, B., Frey, A., Seeber, S., Balkenhol, A. & Bernhardt, R. (2016). Adaptive Messung allgemeiner Komptenzen (MaK-adapt). In K. Beck, M. Landenberger & F. Oser (Hrsg.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT* (Wirtschaft - Beruf - Ethik, Bd. 32). Bielefeld: WBV.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Anhang

Abkürzungsverzeichnis

AIC	Akaike information criterion
ASCOT	Technology-based Assessment of Skills and Competences in Vocational education and training
BIC	Bayesian Information Criterion
BME	Bayes Modal Estimator
cAIC	konsistente (Consistent) AIC
CAT	computerisiertes adaptives Testen
CML	bedingte (Conditional) Maximum Likelihood
DIF	Differential Item Functioning
DIPF	Deutschen Institut für Internationale Pädagogische Forschung
EAP	bayesian Expected A Posteriori estimation
FIT	Fixed Item Testing
ICC	Item Characteristic Curves
IPS	Itempoolscore
IPN	Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik
IRT	Item Response Theorie
LLTM	Linear Logistic Test Model
MaK-adapt	Messung allgemeiner Kompetenzen – adaptiv

MAP	bayesian Maximum A Posteriori estimation
MAT	multidimensionales adaptives Testen
MATE	Multidimensional Adaptive Testing Environment
MLE	Maximum Likelihood Estimation
MML	bedingte (Conditional) Maximum Likelihood
MPI	Maximum-Priority-Index
PISA	Programme for International Student Assessment
SuS	Schülerinnen und Schüler
TIMSS	Trends in International Mathematics and Science Study
TOEFL	Test of English as a Foreign Language
UML	unbedingte (Unconditional) Maximum Likelihood
WLE	Weighted maximum Likelihood Estimation

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass mir die geltende Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität Jena vom 06.05.2009 mit den beiden Änderungen vom 17.11.2010 und 19.06.2012 bekannt ist.

Ich habe die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben.

Bei der Auswahl und Auswertung des Materials wurde ich vornehmlich von meinem Betreuer Prof. Dr. Andreas Frey unentgeltlich unterstützt. Bei den Überlegungen zu dem Themenbereich der Positionseffekte erhielt ich zudem unentgeltliche Unterstützung von meinem Kollegen Sebastian Born. Bei der Überarbeitung des Manuskriptes (Rechtschreibung, Grammatik, Ausdruck etc.) wurde ich von Dr. Claudia Kanz und Dr. Christian Spoden unentgeltlich unterstützt. Zudem wurden Teile der Arbeit von den wissenschaftlichen Hilfskräften Marlen Bernhardt und Robby Trinks sowie der studentischen Hilfskraft Daniel Retzmer korrekturgelesen.

Ich habe keine Hilfe eines Promotionsberaters in Anspruch genommen und Dritte haben weder unmittelbar oder mittelbar geldwerte Leistungen von mir für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Dissertation habe ich nicht als Promotionsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Darüber hinaus habe ich keine gleiche oder in wesentlichen Teilen ähnliche Dissertation oder eine andere Abhandlung bei einer anderen Hochschule oder anderen Fakultät als Dissertation eingereicht.

Jena, 16. September 2016

Raphael Bernhardt

Lebenslauf

Name:	Raphael Bernhardt
Geb.:	12. Mai 1985 in Mühlhausen
Familienstand:	verheiratet, zwei Kinder

Aug. 1995 – Jun. 2001	Schulbesuch (Regelschule Struth)
	Realschulabschluss (Note 1,9)
Aug. 2001 – Jun. 2005	Schulbesuch (Berufliches Gymnasium Mühlhausen)
	Abitur (Note 2,6)
Aug. 2001 – Feb. 2005	Schulbesuch (Berufliche Schulen des Unstrut-Hainich Kreises)
	Abschluss als Technischer Assistent für Informatik (Note 1,9)
Apr. 2005 – Mär. 2011	Studium, Soziologie (Nebenfächer Erziehungswissenschaft und Informatik), FSU Jena
31. Mär. 2011	Hochschulabschluss: Magister Artium (Note: 1,55)
Aug. 2007 – Apr. 2009	berufliche Tätigkeit als studentischer Mitarbeiter, am Institut für Soziologie an der FSU Jena
Okt. 2009 – Mär. 2011	berufliche Tätigkeit als studentischer Mitarbeiter, am Institut für Erziehungswissenschaft an der FSU Jena

Mai 2011 – Apr. 2015	berufliche Tätigkeit als wissenschaftlicher Mitarbeiter, am Institut für Erziehungswissenschaft, an der Professur für empirische Methoden der erziehungswissenschaftlichen Forschung an der FSU Jena
seit Apr. 2015	berufliche Tätigkeit als wissenschaftlicher Mitarbeiter, am Institut für Psychologie, am Lehrstuhl für Methodenlehre und Evaluationsforschung an der FSU Jena

Jena, 16. September 2016

Raphael Bernhardt